

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:

Hiroshi TSUDA

Serial No.: To Be Assigned

Filed: March 12, 1999

For: DOCUMENT FILE GROUP ORGANIZING APPARATUS AND METHOD
THEREOF

Group Art Unit: To Be Assigned

Examiner: To Be Assigned



**SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN
APPLICATION IN ACCORDANCE
WITH THE REQUIREMENTS OF 37 C.F.R. § 1.55**

*Assistant Commissioner for Patents
Washington, D.C. 20231*

Sir:

In accordance with the provisions of 37 C.F.R. § 1.55, Applicants submit herewith a certified copy of each of the following foreign application:

Japanese Appln. No. 10-176749, filed June 24, 1998.

It is respectfully requested that Applicants be given the benefit of the earlier foreign filing date, as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. § 119.

Respectfully submitted,
STAAS & HALSEY

Dated: March 12, 1999

By:

James D. Halsey, Jr.
Registration No. 22,729

700 Eleventh Street, N.W.
Suite 500
Washington, D.C. 20001
(202) 434-1500

PATENT OFFICE
JAPANESE GOVERNMENT



This is to certify that the annexed is a true copy of the
following application as filed with this Office.

Date of Application: June 24, 1998

Application Number: Patent Application
No. 10-176749

Applicant(s): FUJITSU LIMITED

October 9, 1998

Commissioner,
Patent Office Takeshi Isayama

Certificate No. 10-3081505

日 本 国 特 許 庁

PATENT OFFICE
JAPANESE GOVERNMENT

JCS30 U.S. PTO
09/266863
03/12/99

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日

Date of Application:

1998年 6月24日

出 願 番 号

Application Number:

平成10年特許願第176749号

出 願 人

Applicant (s):

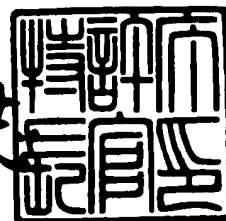
富士通株式会社

CERTIFIED COPY OF
PRIORITY DOCUMENT

1998年10月 9日

特許庁長官
Commissioner,
Patent Office

伴佐山 建志



出証番号 出証特平10-3081505

【書類名】 特許願

【整理番号】 9800955

【提出日】 平成10年 6月24日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 17/30

【発明の名称】 文書整理装置および方法

【請求項の数】 17

【発明者】

 【住所又は居所】 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内

 【氏名】 津田 宏

【特許出願人】

 【識別番号】 000005223

 【氏名又は名称】 富士通株式会社

【代理人】

 【識別番号】 100074099

 【郵便番号】 102

 【住所又は居所】 東京都千代田区二番町8番地20 二番町ビル3F

 【弁理士】

 【氏名又は名称】 大菅 義之

 【電話番号】 03-3238-0031

【選任した代理人】

 【識別番号】 100067987

 【郵便番号】 222

 【住所又は居所】 神奈川県横浜市港北区太尾町1418-305 (大倉山二番館)

 【弁理士】

 【氏名又は名称】 久木元 彰

 【電話番号】 045-545-9280

【手数料の表示】

【予納台帳番号】 012542

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9705047

【プールの要否】 要

【書類名】 明細書

【発明の名称】 文書整理装置および方法

【特許請求の範囲】

【請求項 1】 文書群をキーワードに基づいて整理する文書整理装置であって、

与えられたキーワードからキーワード間の階層関係と連想関係を抽出する関係抽出手段と、

前記階層関係と連想関係をリンクとして用いて、前記文書群にアクセスするためのディレクトリ情報を生成する生成手段と、

前記ディレクトリ情報を出力する出力手段とを備えることを特徴とする文書整理装置。

【請求項 2】 前記関係抽出手段は、前記与えられたキーワードからキーワード対を含む相関ルールを抽出するルール抽出手段と、該相関ルールを評価して、前記階層関係、前記連想関係、および同値関係のいずれかを該キーワード対に付与するルール評価手段を含むことを特徴とする請求項 1 記載の文書整理装置。

【請求項 3】 前記ルール抽出手段は、共起出現頻度の高いキーワード対を前記相関ルールとして抽出し、前記ルール評価手段は、抽出されたキーワード対の関係を該共起出現頻度に基づいて付与することを特徴とする請求項 2 記載の文書整理装置。

【請求項 4】 前記関係抽出手段は、前記与えられたキーワードから共起出現頻度の高いキーワード対を抽出する手段と、該キーワード対の共起出現頻度に基づいて、前記階層関係、前記連想関係、および同値関係のいずれかの関係を該キーワード対に付与する手段を含むことを特徴とする請求項 1 記載の文書整理装置。

【請求項 5】 前記関係抽出手段は、前記与えられたキーワードから部分文字列を切り出し、該与えられたキーワードと部分文字列の包含関係を前記階層関係として抽出する切り出し手段を含むことを特徴とする請求項 1 記載の文書整理装置。

【請求項 6】 前記関係抽出手段は、キーワード対の共起出現頻度に基づい

て第1の階層関係を抽出する手段と、前記与えられたキーワードと部分文字列の包含関係から第2の階層関係を抽出する手段と、該第1および第2の階層関係と外部から与えられた階層関係とをマージする手段とを含むことを特徴とする請求項1記載の文書整理装置。

【請求項7】 前記生成手段は、前記階層関係と連想関係を用いて、トップカテゴリからディレクトリへのパス、該ディレクトリの上位関連語、該ディレクトリのサブカテゴリ、該ディレクトリの連想語、および50音・アルファベット順索引のうち、少なくとも1つ以上のリンクを含むハイパーテキストによる索引を前記ディレクトリ情報として生成し、前記出力手段は、前記文書群を該索引により整理して出力することを特徴とする請求項1記載の文書整理装置。

【請求項8】 前記生成手段は、前記階層関係と連想関係を用いて、トップカテゴリのキーワードから各キーワードへのパスを自動的に計算し、得られたパスを前記トップカテゴリからディレクトリへのパスとして設定することを特徴とする請求項7記載の文書整理装置。

【請求項9】 外部から与えられた同意語リストに基づいて、キーワード間の同値関係を付加する手段をさらに備え、前記生成手段は、該同値関係を含むディレクトリ情報を生成することを特徴とする請求項1記載の文書整理装置。

【請求項10】 外部から与えられた不要語リストに基づいて、キーワードを削除する手段をさらに備え、前記生成手段は、削除されたキーワードを除くディレクトリ情報を生成することを特徴とする請求項1記載の文書整理装置。

【請求項11】 外部から与えられたキーワード間の階層関係を入力する手段をさらに備え、前記生成手段は、該与えられたキーワード間の階層関係を用いてディレクトリ情報を生成することを特徴とする請求項1記載の文書整理装置。

【請求項12】 前記与えられたキーワードを旧キーワードと比較して、新規キーワードを同定する手段をさらに備え、前記出力手段は、該新規キーワードを強調して出力することを特徴とする請求項1記載の文書整理装置。

【請求項13】 前記ディレクトリ情報にアクセスするアクセス手段をさらに備え、利用者は、該ディレクトリ情報を介して前記文書群にアクセスすることを特徴とする請求項1記載の文書整理装置。

【請求項 14】 前記ディレクトリ情報に含まれるキーワードを検索するキーワード検索手段と、前記文書群の文書の内容を検索する文書検索手段をさらに備え、利用者は、該キーワード検索手段および文書検索手段を用いて文書情報を取得することを特徴とする請求項 1 記載の文書整理装置。

【請求項 15】 任意の情報をキーワードに基づいて整理する情報整理装置であって、

与えられたキーワードからキーワード間の階層関係と連想関係を抽出する関係抽出手段と、

前記階層関係と連想関係をリンクとして用いて、前記任意の情報にアクセスするためのディレクトリ情報を生成する生成手段と、

前記ディレクトリ情報を出力する出力手段と
を備えることを特徴とする情報整理装置。

【請求項 16】 文書群をキーワードに基づいて整理するコンピュータのためのプログラムを記録した記録媒体であって、

与えられたキーワードからキーワード間の階層関係と連想関係を抽出するステップと、

前記階層関係と連想関係をリンクとして用いて、前記文書群にアクセスするためのディレクトリ情報を生成するステップと

を含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項 17】 与えられたキーワードからキーワード間の階層関係と連想関係を抽出し、

前記階層関係と連想関係をリンクとして用いて、文書群にアクセスするためのディレクトリ情報を生成し、

前記ディレクトリ情報に基づいて前記文書群を整理することを特徴とする文書整理方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、情報処理装置に蓄えられた大量の文書ファイル群を、その内容に基づいて整理する文書整理装置およびその方法に関する。

【0002】

【従来の技術】

今日、コンピュータネットワークの発達により、大量のオンライン文書情報が溢れてきており、文書検索および文書整理に対する利用者の期待も大きくなっている。例えば、インターネットのホームページ検索サービスでは、大別して次の2種類のサービスまたはそれらの組合せが提供されている。

(a) ディレクトリサービス系

ホームページを階層的に分類して整理する。

(b) 全文検索系

ロボット（検索プログラム）が集めたページの全文検索を行う。

【0003】

ある有名なディレクトリサービスでは、ディレクトリの作成にあたって、次のような方法を取っている。

1. ホームページ作成者が、ホームページを登録したいURL (uniform resource locator) を申請する。

【0004】

2. サービス提供者が、ホームページを階層カテゴリに分類して登録する。

3. 階層カテゴリはサービス提供者独自のものであり、常に変化する。また、1つのホームページは複数のカテゴリに分類される。

【0005】

このサービスでは、サーファと呼ばれる十数名の専門家がディレクトリの作成や情報のメンテナンスを行っており、これにより、常に良質な情報を提供することができる。しかし、大量の文書を分類するための人手をコンスタントに確保するのは、実際には難しい。また、個人が受信した大量の電子メール等を分類するような場合、手作業でディレクトリの作成等を行うのは非常に煩雑である。そこで、コンピュータによる文書自動分類の研究が期待されている。

【0006】

従来、分類学（タクソノミ）では、情報が木構造により分類され、木の分岐において、各子ノードは互いに独立である。また、交叉分類は許されず、情報の配置位置は木構造上の1箇所に限られるという特徴がある。

【0007】

このようなタクソノミの方法を文書検索に用いる場合、文書が木構造により分類され、1つの文書へのパスはただ一つに限られる。しかし、利用者が分類者と同じような分類基準を持っているとは限らないので、文書に辿りつくのが困難な場合があり、必ずしも有効な方法とは言えない。

【0008】

そこで、分類を文書検索に利用する場合には、1つの文書に複数のカテゴリを与えることが考えられる。例えば、インターネットのディレクトリサービスのディレクトリ構造はそうになっている。また、「文書情報分類方法および文書情報分類装置」（特開平8-153121）では、文書群のキーワードから階層的なカテゴリを作成し、文書を複数のカテゴリに格納している。

【0009】

ところで、文書自動分類の研究では、大別して次の2つのアプローチがある。どちらも、利点や欠点があり用途によって使い分けたり、両者を組み合わせたりする必要がある。

（a）クラスタリング

キーワードの統計／表層上の関係をベースにして、与えられた文書群をいくつかの適当なクラスに分割する。このアプローチの利点は、既存の分類にとらわれずに、元の文書群の特徴を反映した分類結果が得られることであり、欠点は、自動化の精度が低いことである。

（b）カテゴリゼーション

与えられた文書が既存の分類におけるどのカテゴリに当てはまるかを判別する。既存の分類としてはシソーラス（概念階層辞書）等があり、文書中のキーワードの分布等を手掛かりにして、その文書を近いカテゴリに当てはめていく。このアプローチの利点は、自動化の精度がクラスタリングより高いことであり、欠点は、分類結果が汎用的で、元の文書群の特徴を反映しないことである。

【0010】

インターネットのディレクトリサービスの多くでは、通常、既存の分類に対してカテゴリゼーションを人手で行う。そして、1つのクラスが大きくなった場合には、クラスを分割するクラスタリング操作を、やはり人手で行っている。

【0011】

例えば、上述の特開平8-153121のシステムは、文書に付加されたキーワードに基づいてクラスタリングを行うシステムである。また、キーワードによるクラスタリングの欠点を補うため、汎用のシソーラスを用い、意味属性の統計を利用して分類精度を上げる研究も報告されている（河合 敦夫，意味属性の学習結果にもとづく文書自動分類方式，情報処理学会論文誌，Vol.33，No.9，pp.1114-1122，1992.）。

【0012】

【発明が解決しようとする課題】

しかしながら、上述した従来の文書分類システムには、次のような問題がある。

【0013】

まず、人手による分類は、ディレクトリを作成したり、運用したりする専門家を必要とし、一般の利用者が分類を行うことは難しい。また、ディレクトリのハイパーテキストのメンテナンスを人手で行う場合、管理者の手間が大きく、単純なミスも発生しやすい。

【0014】

また、タクソノミに従って文書を自動分類すると、通常、情報は1つのカテゴリにしか分類されない。この場合、利用者が分類者と同じような分類基準を持っていなければ、情報に辿りつくのが困難となることがある。また、クラスタリング、カテゴリゼーションともに完全な自動化は不可能であり、分類にゴミやもれがあると、利用者が情報に辿りつくのはさらに困難となる。

【0015】

さらに、上述の河合の報告によれば、クラスタリングの精度は60%ほどで、実用化にはほど遠い。また、カテゴリゼーションは、汎用的な分類にとどまり

、元の文書群の特徴を反映しない。

【0016】

本発明の課題は、情報処理装置に蓄えられた大量の文書群を、その特徴に従って高い精度で自動的に分類する文書整理装置およびその方法を提供することである。

【0017】

【課題を解決するための手段】

図1は、本発明の文書整理装置の原理図である。図1の文書整理装置は、関係抽出手段1、生成手段2、および出力手段3を備え、文書群をキーワードに基づいて整理する。

【0018】

関係抽出手段1は、与えられたキーワードからキーワード間の階層関係4（実線）と連想関係5（破線）を抽出する。生成手段2は、階層関係4と連想関係5をリンクとして用いて、上記文書群にアクセスするためのディレクトリ情報を生成する。そして、出力手段3は、そのディレクトリ情報を出力する。

【0019】

関係抽出手段1により抽出される階層関係4は、キーワードの間の概念上の上下関係を表し、連想関係5は、階層関係ほど緊密ではないが、一方のキーワードから他方のキーワードが連想されるような比較的緩やかな関係を表す。この連想関係5により、階層関係4を持たないキーワード同士を関係付けることが可能になる。

【0020】

生成手段2は、階層関係4だけでなく連想関係5もリンクとして用いて、文書に付加されたキーワード間の関係を表すディレクトリ情報を生成し、文書を分類する。そして、出力手段3は、利用者が文書群にアクセスできるように、ディレクトリ情報を文書群の索引として提示する。

【0021】

文書群の分類結果を表すディレクトリ情報に連想関係5をリンクとして付加することにより、階層関係4だけでは得られなかったフレキシブルなアクセスが可

能となる。したがって、文書群の特徴を分類に反映させることがより容易になり、高精度の分類結果が自動的に得られる。

【0022】

例えば、図1の関係抽出手段1は、後述する図2のキーワード関係抽出器42に対応し、生成手段2は、図2のディレクトリファイル生成器43に対応し、出力手段3は、図2の表示装置14およびディレクトリアクセス部44に対応する。

【0023】

【発明の実施の形態】

以下、図面を参照しながら、本発明の実施の形態を詳細に説明する。

本発明においては、管理に要する人手を最小限にするために、管理者は、キーワードの階層関係の極一部だけを辞書や人手で文書整理装置に与える。文書整理装置は、それを元に統計処理および文字列処理を行って、キーワード間の関係を自動的に取り出し、ディレクトリファイルを自動生成する。また、既存の分類方法の精度上の限界を克服するために、文書の分類（階層関係）にはこだわらず、利用者が文書にアプローチするための多様なリンクを提供することをディレクトリの主眼とする。この文書整理装置の動作の概要は次のようになる。

【0024】

1. まず、入力文書にはその内容を表すキーワードが付与されているものとする。キーワード付与の方法としては、本発明とは独立に、任意の方法を用いることができる。例えば、キーワードを手で付けてもよいし、文書内容からキーワード抽出技術により自動的に取り出してもよい。

【0025】

2. 次に、管理者は、ディレクトリのトップとなるキーワード列、キーワードの同値関係、キーワードの階層（上下）関係、および不要キーワードのリストを、明示的に文書整理装置に与える。

【0026】

3. 文書整理装置は、管理者が明示した不要キーワードのリストと、キーワード変換ルールに基づいて、文書に付与されたキーワードを整える（データクリー

ニング)。

【0027】

4. 文書整理装置は、各キーワードを含む文書集合を計算し、管理者が同意語として明示した同値関係にある2つ以上のキーワードの文書集合をマージする。

5. 文書整理装置は、任意の2つのキーワードに対して、それらの文書集合の関係をもとに、キーワード間の関連性（同値関係／階層関係／連想関係）を計算する。

【0028】

6. 文書整理装置は、管理者が明示した階層関係および計算により得られた階層関係を元に、管理者が与えたトップキーワードから特定のキーワードまでの最短距離のキーワード列を、そのキーワード（およびそれを有する文書）へのパスとして計算する。

【0029】

7. 文書整理装置は、まだパスの付いていないキーワードについて、そのキーワードから最も少ない上位関連語列により辿りつけるパスの付いたキーワードを探索する。そして、得られたキーワードのパスに上位関連語列を付加したものを、元のキーワードのパスとする。

【0030】

8. それでもパスが付けられないキーワードについては、文書整理装置は、トップキーワードに“その他”を設け、その下位語として登録する。

9. 文書整理装置は、単語読み付加器を用いて、キーワードに読みを与える。

【0031】

10. 文書整理装置は、管理者が明示したキーワード間の階層関係、自動的に得られたキーワード間の関係、自動的に計算されたパス、キーワードの読み、およびキーワード分割装置（部分文字列切り出し器）を用いて、ディレクトリのハイパーテキストを作成する。

【0032】

11. 利用者は、作成されたディレクトリを文書群の索引として用いて、次のような操作を行う。

- ・トップレベルから下位関連語を用いて文書の検索範囲を絞り込む。

【0033】

- ・上位関連語を用いて検索範囲を広げる。
- ・パスで現在位置を確認し、上位に飛ぶ。
- ・連想関係のキーワードを用いて別の概念に飛ぶ。

【0034】

- ・50音・アルファベット順索引を用いてキーワードを探す。

図2は、このような文書整理装置の構成図である。図2の文書整理装置は、処理装置11、二次記憶装置12、入力装置13、および表示装置14を備える。処理装置11は、例えば、CPU（中央処理装置）とメモリを含み、入力装置13は、キーボード、マウス等に対応し、表示装置14は、ディスプレイ等に対応する。

【0035】

また、二次記憶装置12は、電子化された文書群のデータ21と、ディレクトリ管理者が辞書等を用いて作成／管理する管理ファイル22を格納する。データ21の各文書には、それぞれ、複数のキーワードが付加されている。また、管理ファイル22には、同意語および不要語の集合31と、キーワードの概念上の階層関係を表すデータ32と、ディレクトリのトップとなるキーワードの集合33が含まれている。

【0036】

ここで、同意語とは、“コンピュータ”と“計算機”のように、互いに同意語関係にあるキーワードの集合を表し、不要語とは、差別用語のように、ディレクトリに用いたくないキーワードの集合を表す。また、階層関係としては、例えば、“コンピュータ”を上位語（上位関連語）として、“ハードウェア”、“ソフトウェア”等の下位語（下位関連語）を定義する情報が用いられる。

【0037】

また、処理装置11は、キーワード整形器41、キーワード関係抽出器42、ディレクトリファイル生成器43、ディレクトリアクセス部44、および検索部45を含む。これらは、例えば、プログラムにより記述されたソフトウェアコン

ポーネントに対応し、処理装置 11 の特定のプログラムコードセグメントに格納される。

【0038】

キーワード整形器 41 は、いわゆるデータクリーニングを行い、文書単語対 51 と文書メタ情報 52 を出力する。ここでは、文書データ 21 の中のゴミを取ったり、キーワードを統一したりする処理が行われる。文書単語対 51 は、整形後の文書とキーワードの対であり、文書メタ情報 102 は、文書の諸々の情報（URL、タイトル等）やキーワードの読み等の情報である。

【0039】

キーワード関係抽出器 42 は、データ 32、51 から、キーワード間の階層関係 53、同値関係 54、および連想関係 55 を計算する。これらの関係は、いずれも 2 つ以上のキーワード間になんらかの関連性があることを表している。

【0040】

階層関係 53 は、キーワードの間の概念上の上下関係を表すデータであり、同値関係 54 は、キーワードの間の概念上の同位関係を表すデータである。例えば、同意語の集合に含まれるキーワードは、互いに同値関係にある。また、連想関係 55 は、階層関係も同値関係も持たないキーワード間において、一方のキーワードから他方のキーワードが連想されるような関係を表すデータである。

【0041】

ディレクトリファイル生成器 43 は、データ 33、52、53、54、55 から、ハイパーテキスト形式のディレクトリファイル 56 を生成し、ディレクトリアクセス部 44 は、入力装置 13 から入力される利用者の指示に従って、ディレクトリファイル 56 の内容を表示装置 14 に表示したり、検索部 45 を起動したりする。

【0042】

検索部 45 は、利用者の指示に従って、ディレクトリファイル 56 や文書データ 21 の内容を全文検索する。文書データ 21 の全文検索では、文書の本文に含まれる任意の文字／単語列の検索が行われる。

【0043】

次に、図3から図5までを参照しながら、各種データのデータ構造について説明する。

図2の管理ファイル22に含まれるデータ31、32、33は、例えば、図3に示すような3つのテーブルに格納される。図3の同意語／不要語テーブル31は、[代表語ID, 同意語または不要語のIDの集合]の組合せを表す。例えば、代表語ID=005のキーワードは、ID=010, 021, 022の他のキーワードと同意語の関係にあり、代表語IDが空(nil)のID集合(ID=077, 082)は、不要語である。

【0044】

また、階層関係テーブル32は、[上位語ID, 下位語IDの集合]の組合せを表す。例えば、上位語ID=002のキーワードは、階層関係における下位語としてID=004, 008のキーワードを有する。また、トップキーワードIDリスト33は、トップディレクトリとして用いられるキーワードのIDの列(ID=001, 008, ...)を表す。

【0045】

また、図2の文書単語対51、階層関係53、同値関係54、連想関係55は、例えば、図4に示すような形式の2つのテーブルに格納される。図4の文書単語対テーブル61は、

[文書ID, キーワードIDの集合]

の組合せを格納し、キーワードテーブル62は、

[キーワードID(KID), キーワード, 読み, 上位語IDの集合(UP), 下位語IDの集合(DOWN), 連想語IDの集合(Rel), 同値関係にあるキーワードIDの集合(Eq), パス, 新規語フラグ(new)]

の組合せを格納する。

【0046】

ここで、上位語IDの集合と下位語IDの集合は階層関係53に対応し、連想語IDの集合は連想関係55に対応し、同値関係にあるキーワードIDの集合は同値関係54に対応する。また、新規語フラグは、キーワードが新規語に対応する場合は“1”となり、そうでない場合は“0”となる。

【0047】

例えば、文書単語対テーブル61の文書ID=000の文書には、ID=000, 081のキーワードが付加されており、ID=000のキーワードは、キーワードテーブル62の“スポーツ”に相当する。

【0048】

このキーワードの読みとしては、“すぽーつ”が登録されており、その上位語としては、ID=008, 022の2つのキーワードが登録されている。また、その下位語としては、ID=025のキーワードが登録されており、連想語としては、ID=038, 087の2つのキーワードが登録されている。この段階では、同値関係にあるキーワードとパスは登録されておらず、新規語フラグ“1”は、“スポーツ”が新規語であることを表している。

【0049】

また、図2の文書メタ情報52は、図5に示すような文書情報テーブルの形式で格納される。この文書情報テーブル52は、

【文書ID, タイトル, 説明, 更新日時, 一次情報へのリンク】
の組合せを格納する。

【0050】

ここで、説明は、文書の要約または最初の数行の記述に対応し、更新日時は、文書が最後に更新された日時を表し、一次情報へのリンクは、インターネット上の対応するURL等を表す。

【0051】

例えば、文書ID=001の文書のタイトルは“ニュースの読み方”であり、その説明は“comp,fj をmuleから...”であり、更新日時は“1998年2月10日15時38分”であり、一次情報へのリンクは“http://www.xxxx(URL)”である。

【0052】

次に、図6および図7を参照しながらキーワード整形器41の処理を詳細に説明する。

図6は、キーワード整形器41の構成図である。図6のキーワード整形器41

は、サブコンポーネントとして、キーワード統一器 7 1 および単語読み付加器 7 2 を含む。キーワード統一器 7 1 は、与えられたキーワードの文字コードを統一し、単語読み付加器 7 2 は、与えられた単語の読みを生成する。

【0053】

このキーワード整形器 4 1 は、図 7 に示すような処理を行って、文書単語対 5 1 と文書情報テーブル 5 2 を生成する。文書単語対 5 1 は、図 4 に示したように、文書単語対テーブル 6 1 とキーワードテーブル 6 2 から成っている。

【0054】

まず、キーワード統一器 7 1 は、文書データ 2 1 の文書に付加されたすべてのキーワードについて、文字を統一する（ステップ S 1 0）。ここでは、キーワードに含まれる漢字コードを、例えば、EUCコードのような特定のコードに統一したり、半角のカタカナ／英数字を全角のカタカナ／英数字に統一したり、キーワード中の記号や空白を除去したりするような、文字単位でのキーワード統一処理が行われる。

【0055】

次に、キーワード整形器 4 1 は、1 つの文書のキーワード情報のうち、キーワード ID とキーワードをキーワードテーブル 6 2 に追加し、読みと新規語フラグを付加する（ステップ S 1 1）。

【0056】

このとき、日本語のキーワードの読みについては、単語読み付加器 7 2 が生成した読みを設定し、英語のキーワードの読みについては、キーワード自身と同一に設定する。また、追加したキーワードが旧キーワードの集合 7 3 に登録されていれば、新規語フラグを“0”に設定し、旧キーワードの集合 7 3 に登録されていなければ、新規語フラグを“1”に設定する。旧キーワードの集合 7 3 は、前回のディレクトリ生成時に生成されたキーワードテーブル 6 2 のキーワードの集合を表す。

【0057】

次に、キーワード整形器 4 1 は、文書 ID およびキーワード ID を、文書メタ情報 5 2 およびキーワードテーブル 6 2 へのリンクとして文書単語対テーブル 6

1に追加する（ステップS12）。このとき、対応する文書に関するデータを文書情報テーブル52に追加しておく。

【0058】

次に、文書データ21の中に未処理の文書があるかどうかを判定し（ステップS13）、そのような文書があれば、ステップS11以降の処理を繰り返す。そして、すべての文書について登録が完了すると、次に、同意語／不要語テーブル31の情報をキーワードテーブル62に追加する（ステップS14）。

【0059】

ここでは、テーブル31の代表語IDに対応するキーワードのフィールドEqに、その同意語として定義されたIDを追加し、不要語として定義されたIDをキーワードテーブル62のエントリからすべて消去する。また、同時に、文書単語対テーブル61から不要語のIDをすべて消去する。これにより、管理者が指定した同意語がキーワードテーブル62に登録され、不要語が文書単語対テーブル61およびキーワードテーブル62から消去される。

【0060】

次に、キーワードテーブル62のエントリとして、“コンピューター”と“コンピュータ”のように、語尾の文字“ー”が有るキーワードとそれが無いキーワードとが併存する場合には、それらを統一する（ステップS15）。ここでは、例えば、“コンピューター”のフィールドEqに“コンピュータ”のキーワードIDを登録することにより、これらの2つのキーワードを縮退させることができる。

【0061】

また、“コンピュータグラフィックス”と“コンピュータ”のように、あるキーワードとそれに含まれる他のキーワードに関しても統一し、それらのエントリが併存する場合には、一方のフィールドEqに他方のキーワードIDを登録する。

【0062】

次に、こうして生成されたキーワードテーブル62のキーワードのリストを、旧キーワードの集合73として保存し（ステップS16）、処理を終了する。旧

キーワードの集合 73 は、次回のディレクトリ生成時に、前述のステップ S11 の処理で参照される。

【0063】

次に、図 8 から図 15 までを参照しながらキーワード関係抽出器 42 の処理を詳細に説明する。

図 8 は、キーワード関係抽出器 42 の構成図である。図 8 のキーワード関係抽出器 42 は、部分文字列切り出し器 81、相関ルール抽出器 82、ルール評価器 83、およびマージャ 84 を含み、文書単語対 51（図 4 の文書単語対テーブル 61 とキーワードテーブル 62 に対応）と管理者が与えた階層関係 32 から、階層関係 53、同値関係 54、および連想関係 55 を生成する。

【0064】

部分文字列切り出し器 81 は、複合語句のキーワードを適当な文字列に分割し、部分語関係のデータ 85 を生成する。データ 85 は、“情報検索”と“情報”／“検索”のように、キーワード間で一方が他方を含むような関係を表す。

【0065】

相関ルール抽出器 82 は、キーワードの出現頻度から、キーワード間の関係を表す相関ルール 86 を抽出する。ルール評価器 83 は、相関ルール 86 を評価し、それを階層関係、同値関係 54、および連想関係 55 に分割する。マージャ 84 は、階層関係 32、部分語関係 85、およびルール評価器 83 が生成した階層関係を合わせて、階層関係 53 を生成する。

【0066】

図 9 は、キーワード関係抽出器 42 の処理のフローチャートである。まず、相関ルール抽出器 82 は、文書単語対テーブル 61 とキーワードテーブル 62 から相関ルール 86 を抽出し、それを相関ルールの集合 AR とする（ステップ S21）。

【0067】

次に、ルール評価器 83 は、AR に基づいてルール評価を行い、それを上位語関係 up1、下位語関係 down1、同値関係 eq1、および連想関係 rw1 に分割する（ステップ S22）。2つのキーワード w1、w2 について、up1 お

よび $down1$ は、“ $w1 > w2$ ” ($w1$ は $w2$ の上位語、または $w2$ は $w1$ の下位語) のように記述され、 $eq1$ は“ $w1 = w2$ ” のように記述され、 $rw1$ は“ $w1 \sim w2$ ” のように記述される。

【0068】

次に、ルール評価器 83 は、同値関係 $eq1$ をキーワードテーブル 62 のフィールド E_q に登録する (ステップ S23)。 $eq1$ は、“ $w1 = w2$ ” のようなキーワード間の関係の集合である。そこで、キーワード $w1$ のフィールド E_q のエントリに、キーワード $w2$ の ID を追加する。

【0069】

次に、ルール評価器 83 は、連想関係 $rw1$ をキーワードテーブル 62 のフィールド R_{e1} に登録する (ステップ S24)。 $rw1$ は、“ $w1 \sim w2$ ” のようなキーワード間の関係の集合である。そこで、キーワード $w1$ のフィールド R_{e1} のエントリに、キーワード $w2$ の ID を追加する。

【0070】

次に、部分文字列切り出し器 81 は、キーワードテーブル 62 のキーワードの中から部分文字列を取り出し、それらの包含関係を表す部分語関係 85 を生成する (ステップ S25)。そして、それを部分語関係の集合 $sr1$ とおく。 $sr1$ には、例えば、“コンピュータ”と“グラフィックス”が“コンピュータグラフィックス”の部分文字列であることが登録される。

【0071】

次に、マージャ 84 は、部分語関係 $sr1$ を参照し、キーワード kw が部分文字列 w_1, w_2, \dots, w_n を含んでいる場合、 w_i が kw の上位語 ($i = 1, 2, \dots, n$) であるという階層関係に変換し、それを“ $w_i > kw$ ”と記述する (ステップ S26)。こうして得られた階層関係の集合を $sr2$ とする。

【0072】

次に、マージャ 84 は、管理者が定義した階層関係 32 に $sr2$ をマージし、その結果得られた階層関係を HR とする (ステップ S27)。ここで、ある階層関係 $S1$ に他の階層関係 $S2$ をマージするとは、 $S2$ の要素のうち、 $S1$ の各要素と矛盾しないものを新たに $S1$ に加える操作を意味する。例えば、 $S2$ の要素

“ $w_1 > w_2$ ”がS1になく、“ $w_2 > w_1$ ”という要素もS1にないとき、“ $w_1 > w_2$ ”がS1に加えられる。

【0073】

ここでは、階層関係32の方がs r 2より優先されるため、s r 2の要素のうち、階層関係32の要素と矛盾するものは、階層関係32に加えられない。例えば、階層関係32が、

{コンピュータ>ソフトウェア, コンピュータ>ハードウェア, ソフトウェア>グループウェア, グローバルネットワーク>ネットワーク}

であり、s r 2が、

{コンピュータ>パーソナルコンピュータ, ネットワーク>グローバルネットワーク}

である場合、階層関係32にs r 2をマージした結果は、

{コンピュータ>ソフトウェア, コンピュータ>ハードウェア, コンピュータ>パーソナルコンピュータ, ソフトウェア>グループウェア, グローバルネットワーク>ネットワーク}

となる。

【0074】

次に、マージャ84は、ステップS22で得られたdown1をHRにマージし、その結果を改めてHRとする(ステップS28)。これにより、文字列の包含関係により得られた階層関係s r 2が、ルール評価により得られた階層関係down1よりも優先されることになる。

【0075】

次に、マージャ84は、HRの各要素“ $w_1 > w_2$ ”について、キーワードテーブル6.2のキーワードw1のレコードのフィールドDOWNに、キーワードw2のIDを追加する(ステップS29)。

【0076】

次に、マージャ84は、ステップS22で得られたup1をHRにマージし、その結果を改めてHRとする(ステップS30)。これにより、上位語関係もHRに含まれる。

【0077】

次に、マージャ84は、HRの各要素“ $w_1 > w_2$ ”について、キーワードテーブル62のキーワード w_2 のレコードのフィールドUPに、キーワード w_1 のIDを追加して（ステップS31）、処理を終了する。

【0078】

図10は、図9のステップS21で行われる相関ルール抽出処理と、ステップS22で行われるルール評価処理のフローチャートである。ここで、与えられる入力データは、キーワード整形済みの文書単語対テーブル61とキーワードテーブル62である。

【0079】

まず、相関ルール抽出器82は、キーワードテーブル62の同値関係を参照して、文書毎に同値関係を解消したキーワードの集合を生成し、データマイニングの相関ルール抽出アルゴリズムを利用して、相関ルールを抽出する（ステップS41）。

【0080】

ここで、同値関係を解消したキーワードの集合とは、文書単語対テーブル61に文書と対応して登録されているキーワードに、それらのキーワードのフィールドEqに登録されたキーワードを加え、相異なるものだけを残したキーワードの集合を意味する。

【0081】

また、相関ルール抽出アルゴリズムによれば、キーワード対（H，B）に関する統計情報に基づき、文書とキーワード集合の対から $H \rightarrow B$ という形式のルールが抽出される。ここで、 $H \rightarrow B$ というルールは、キーワード対（H，B）の共起出現頻度を表すサポート $\text{sup}(H \rightarrow B)$ および確信度 $\text{conf}(H \rightarrow B)$ により特徴付けられる。これらの値は、次式により定義される。

$$\begin{aligned} & \text{sup}(H \rightarrow B) \\ &= (\text{H and Bを有する文書数}) / \text{全文書数} \\ & \text{conf}(H \rightarrow B) \\ &= (\text{H and Bを有する文書数}) / (\text{Hを有する文書数}) \end{aligned}$$

ステップS41では、SupおよびCfを適当な閾値として、すべてのキーワードH、Bの組合せのうち、 $\text{sup}(H \rightarrow B) \geq \text{Sup}$ および $\text{conf}(H \rightarrow B) \geq \text{Cf}$ を満たすような、HとBの組合せをルールとして取り出す。

【0082】

次に、ルール評価器83は、このようにして取り出されたルールの集合を分割して、キーワード間の階層関係、同値関係、および連想関係を取り出す。これにより、各ルールに含まれるキーワード対の関係が自動推定される。

【0083】

ここでは、kwと w_i (w_i)をルールに含まれるキーワードIDとして、 $\text{conf}(kw \rightarrow w_i)$ の値をX軸に取り、 $\text{conf}(w_i \rightarrow kw)$ の値をY軸に取って、図11に示すような確信度のXY平面を考える。

【0084】

このとき、ルールの集合に含まれるkwと w_i のすべての組合せについて、 $\text{Cf} \leq \text{conf}(kw \rightarrow w_i) \leq 1$ 、 $\text{Cf} \leq \text{conf}(w_i \rightarrow kw) \leq 1$ が成り立つ。したがって、各 w_i について、点 $(X, Y) = (\text{conf}(kw \rightarrow w_i), \text{conf}(w_i \rightarrow kw))$ をXY平面上にプロットすると、×印で示されるように、X軸、Y軸、直線 $X = \text{Cf}$ 、直線 $X = 1$ 、直線 $Y = \text{Cf}$ 、および直線 $Y = 1$ で囲まれた領域内に点が分布することになる。

【0085】

そこで、適当な閾値をTx、Tyとして、直線 $X = \text{Tx}$ と $Y = \text{Ty}$ によりこの領域を4つの部分領域91、92、93、94に分割し、それぞれの部分領域に含まれる w_i に、kwとの間の階層関係、同値関係、および連想関係のいずれかを付与する。

ルール評価器83は、まず、変数kwを000（キーワードIDの最小値）とおく（ステップS42）。そして、kwを左辺または右辺に持つルールを要素とする集合をSとおき、変数Tx、Tyの初期値をそれぞれ1に設定する（ステップS43）。

【0086】

次に、Sの要素のうち、 $\text{conf}(kw \rightarrow w_i) > \text{Tx}$ となるようなルールに

含まれる w_i の数 $s_x(S, T_x)$ を求め、それが一定値 \min_x (\min_x) を越えるまで、 T_x の値を徐々に下げていく (ステップ S44~S47)。この処理により、図 11 の領域 91、92 を合わせた領域のキーワード数が \min_x 以上になるように、 T_x が決められる。ただし、 T_x は最小確信度 Cf より下には下げないものとする。

【0087】

ここでは、まず、 $s_x(S, T_x)$ を \min_x と比較し (ステップ S44)、 $s_x(S, T_x)$ が \min_x 以下であれば、 T_x の値を 0.1 だけ下げる (ステップ S45)。そして、 T_x を Cf と比較し (ステップ S46)、 T_x が Cf より大きければ、ステップ S44 以降の処理を繰り返す。そして、ステップ S44 において $s_x(S, T_x)$ が \min_x を越えれば、ステップ S48 以降の処理に移る。また、ステップ S46 において T_x が Cf 以下となった場合は、 $T_x = Cf$ とおき (ステップ S47)、ステップ S48 以降の処理に移る。

【0088】

次に、 S の要素のうち、 $\text{conf}(w_i \rightarrow kw) > Ty$ となるようなルールに含まれる w_i の数 $s_y(S, Ty)$ を求め、その数が一定値 \min_y (\min_y) を越えるまで、 Ty の値を徐々に下げていく (ステップ S48~S51)。この処理により、図 11 の領域 91、93 を合わせた領域のキーワード数が \min_y 以上になるように、 Ty が決められる。ただし、 Ty は最小確信度 Cf より下には下げないものとする。

【0089】

ここでは、まず、 $s_y(S, Ty)$ を \min_y と比較し (ステップ S48)、 $s_y(S, Ty)$ が \min_y 以下であれば、 Ty の値を 0.1 だけ下げる (ステップ S49)。そして、 Ty を Cf と比較し (ステップ S50)、 Ty が Cf より大きければ、ステップ S48 以降の処理を繰り返す。そして、ステップ S48 において $s_y(S, Ty)$ が \min_y を越えれば、ステップ S52 以降の処理に移る。また、ステップ S50 において Ty が Cf 以下となった場合は、 $Ty = Cf$ とおき (ステップ S51)、ステップ S52 以降の処理に移る。

【0090】

次に、図11に示したように、 $k w$ に関連する確信度を表す点 $(X, Y) = (\text{conf}(k w \rightarrow w_i), \text{conf}(w_i \rightarrow k w))$ を XY 平面上にプロットする(ステップS52)。

【0091】

このとき、右上の矩形領域91に属する各点は、 X 、 Y の値が共に1に近く(大きく)、図12に示すように、 w_i の文書集合(w_i をキーワードとして有する文書の集合)は $k w$ の文書集合とほぼ重なると考えられる。このため、 w_i は $k w$ と同値関係にあるものとみなされる。

【0092】

また、右下の矩形領域92に属する各点は、 X の値が1に近く Y の値が小さいので、図13に示すように、 w_i の文書集合は $k w$ の文書集合をほぼ含んでいると考えられる。このため、 w_i は $k w$ の上位語であるものとみなされる。

【0093】

また、左上の矩形領域93に属する各点は、 Y の値が1に近く X の値が小さいので、図14に示すように、 w_i の文書集合は $k w$ の文書集合にほぼ含まれると考えられる。このため、 w_i は $k w$ の下位語であるものとみなされる。

【0094】

また、左下の領域94に属する各点は、 X 、 Y の値が共に小さいが最小確信度 Cf 以上であるので、 w_i の文書集合と $k w$ の文書集合は、上述のような関係はないが、図15に示すように、なんらかの関連性を持つと考えられる。このため、 w_i は $k w$ の連想語であるものとみなされる。

【0095】

そこで、 S の要素を対応する4つのグループに分割し、領域91の w_i については同値関係“ $k w = w_i$ ”を $eq1$ に加え、領域94の w_i については連想関係“ $k w \sim w_i$ ”を $rw1$ に加える。また、領域92の w_i については上位語関係“ $w_i > k w$ ”を $up1$ に加え、領域93の w_i については、下位語関係“ $k w > w_i$ ”を $down1$ に加える。こうして、 $k w$ に関するすべてのキーワード関係が抽出される。同値関係は、上位語関係および下位語関係とともに、広義の階層関係に属するとも考えられる。

【0096】

次に、ルール評価器 83 は、 kw に 1 を加算して（ステップ S53）、 kw をキーワード ID の最大値 max_{kw} (max_kw) と比較する（ステップ S54）。そして、 kw が max_{kw} を越えていなければ、ステップ S43 以降の処理を繰り返し、 kw が max_{kw} を越えると、処理を終了する。

【0097】

次に、図 16 から図 19 までを参照しながらディレクトリファイル生成器 43 の処理を詳細に説明する。

ディレクトリファイル生成器 43 は、キーワード関係抽出器 42 が生成した階層関係 53、同値関係 54、連想関係 55、管理者が与えたトップキーワード ID リスト 33、およびキーワード整形器 41 が生成した文書メタ情報 52 から、ディレクトリファイル 56 を生成する。

【0098】

ディレクトリファイル 56 は、図 16 に示すような 3 種類のハイパーテキストファイル 101、102、103 から成り、各ファイルの間には互いにリンクが張られている。

【0099】

図 16 において、ディレクトリトップファイル 101 は、ディレクトリの入口に対応するファイルであり、1 つだけ設けられる。このファイル 101 には、キーワード検索の入力窓 104、トップキーワード 105 (KL)、および 50 音・アルファベット順索引 106 が含まれている。トップキーワード 105 の各キーワード KL からは、ディレクトリ中間ファイル 103 へリンクが張られており、50 音・アルファベット順索引 106 の各文字からは、50 音・アルファベット順索引中間ファイル 102 へリンクが張られている。

【0100】

50 音・アルファベット順索引中間ファイル 102 は、“あ行”、“か行”等キーワードの読みにより複数のファイルに分割される。各ファイルのキーワード KL からは、ディレクトリ中間ファイル 103 にリンクが張られている。

【0101】

ディレクトリ中間ファイル103は、キーワード毎に設けられ、ヘッダ107、パス108、上位関連語109、サブカテゴリ110、および文書リスト111の各部分から構成される。

【0102】

ヘッダ107には、ファイル103のタイトルとなるキーワードと、その同意語リストが記述され、パス108には、トップキーワードからそのキーワードまでの経路の1つがキーワード列として記述される。パス108の各キーワードKLからは、そのキーワードのディレクトリ中間ファイル103へリンクが張られている。

【0103】

また、上位関連語109には、上位キーワード列が記述され、サブカテゴリ110には、下位キーワード列が記述される。各キーワードKLからは、そのキーワードのディレクトリ中間ファイル103へリンクが張られている。

【0104】

文書リスト111には、そのキーワードと関連付けられた各文書のタイトルと内容が記述される。タイトルからは、文書の一次情報（URL等）へのリンクが張られている。また、各文書に付加されたキーワードのうち、そのファイルのキーワードと連想関係にあるものが連想語として記述され、各連想語からはそのディレクトリ中間ファイル103へリンクが張られている。検索窓112は、文書の内容を検索する際の入力窓である。

【0105】

なお、図16において、★印の付いたキーワードは新規語に対応する。この印は、図4のキーワードテーブル62のフィールドnewに“1”が設定されたキーワードに対して付加され、それらが新規語であることを強調している。

【0106】

図17は、ディレクトリファイル生成器43の処理のフローチャートである。ディレクトリファイル生成器43は、まず、ディレクトリトップファイル101を生成し、データ33のキーワードを領域105に記述する（ステップS61）。また、50音・アルファベット順索引中間ファイル102へのリンクを領域1

06に記述する。

【0107】

次に、50音・アルファベット順索引中間ファイル102を生成し、キーワードテーブル62に登録された読みを取得して、同一の読みで始まるキーワードを1つのファイル102にまとめる（ステップS62）。

【0108】

次に、トップキーワードから各キーワードまでの最短パスを計算し、それをキーワードテーブル62のパスのフィールドに登録する（ステップS63）。そして、キーワード毎にディレクトリ中間ファイル103を生成して（ステップS64）、処理を終了する。

【0109】

ステップS64では、キーワードテーブル62のフィールドEqに登録されたキーワードを領域107に記述し、計算された最短パスを領域108に記述し、キーワードテーブル62のフィールドUPに登録されたキーワードを領域109に記述し、フィールドDOWNに登録されたキーワードを領域110に記述する。

【0110】

また、文書単語対テーブル61から、ファイル103のタイトルキーワードが付加された文書のIDを取得し、文書情報テーブル52から、そのタイトル、説明、および一次情報へのリンクを取得する。そして、それらを領域111に記述する。このとき、キーワードテーブル62のフィールドRelに登録されたキーワードを、連想語として記述する。したがって、ファイル103の生成において、パス108だけが新規に作成される情報である。

【0111】

ここで、トップキーワードからあるキーワードまでのパスとしては、トップから階層関係だけを辿って到達するものだけでなく、それに連想関係を加えて到達するものも存在する。これは、前述のような階層関係の設定方法では、トップから階層関係だけで必ずしもすべてのキーワードに到達できるという保証がないためである。

【0112】

そもそも、パスはハイパーテキストにおいて利用者が迷子にならないための仕組みである。本実施形態においては、利用者は、パスを逆に辿ることで、あるキーワードからトップキーワードへ到達することができる。

【0113】

図18および図19は、図17のステップS63におけるパス生成処理のフローチャートである。ここでは、最初に階層関係だけでパスの生成を試み、パスが生成されなかったキーワードについては、連想関係も加えてパス生成を試みる。それでもトップと結びつけられないキーワードについては、便宜上、トップの直下に“その他”というカテゴリを生成し、そこに直接結びつける。

【0114】

ディレクトリファイル生成器43は、まず、トップキーワードIDリスト33の各キーワードについて、キーワードテーブル62のパスのフィールドに“top”を登録し（図18、ステップS71）、それらのトップキーワードのリストをS1とする（ステップS72）。

【0115】

次に、階層関係だけを辿ってパスを設定する（ステップS73～S80）。ここでは、幅優先探索を行っており、各時点での最もパスが長いキーワードがS1に入っている。

【0116】

ディレクトリファイル生成器43は、まず、S1が空かどうかを判定し（ステップS73）、それが空でなければ、S1からキーワードwを取り出す（ステップS74）。そして、キーワードテーブル62において、キーワードwのフィールドDOWNに登録されたキーワード集合をS2とする（ステップS75）。

【0117】

次に、S2が空かどうかを判定し（ステップS76）、それが空でなければ、S2からキーワードuを取り出し（ステップS77）、そのパスのフィールドが空かどうかを調べる（ステップS78）。

【0118】

まだ、キーワードuのパスが設定されていなければ、wのパス+wをパスとして設定し、キーワードuをS3に加えて（ステップS79）、ステップS76以降の処理を繰り返す。S3は、パスが設定されたキーワードの集合を表し、最初は空に設定されている。ステップS78においてキーワードuのパスが既に設定されていれば、そのままステップS76以降の処理を繰り返す。

【0119】

そして、ステップS76においてS2が空になると、S3を改めてS1とおき、S3を空に設定して（ステップS80）、ステップS73以降の処理を繰り返す。これにより、キーワードの木構造のトップから下位に向かって、各ノードにパスが設定されていく。

【0120】

ステップS73においてS1が空になると、次に、キーワードテーブル62においてまだパスの設定されていないキーワードに対して、階層関係および連想関係を辿ってパスを設定する（ステップS81～S89）。

【0121】

ディレクトリファイル生成器43は、まず、その時点でまだパスが設定されていないキーワードの集合をS4とし、S6を空に設定する（ステップS81）。次に、S4が空かどうかを判定し（ステップS82）、それが空でなければ、S4からキーワードvを取り出す。そして、キーワードテーブル62において、キーワードvのフィールドUP、DOWN、Relを合わせたキーワードの集合をS5とする（ステップS83）。

【0122】

次に、S5のキーワードでパスが設定されているもののうち、最短パスのキーワードをwとする（ステップS84）。S5のキーワードのパスがすべて空の場合には、wも空となる。

【0123】

次に、キーワードwが空かどうかを判定し（ステップS85）、それが空でなければ、wのパス+wをキーワードvのパスとして設定し、S6にキーワードvを加える（ステップS86）。そして、ステップS82以降の処理を繰り返す。

また、キーワードwが空であれば、そのままステップS 8 2以降の処理を繰り返す。

【0124】

そして、ステップS 8 2においてS 4が空になると、次に、S 6が空かどうかを判定し（ステップS 8 7）、それが空でなければ、ステップS 8 1以降の処理を繰り返す。

【0125】

そして、ステップS 8 7においてS 6が空になると、それ以上処理を繰り返しても新たなパスは設定されないと判断し、その時点でまだパスが設定されていないキーワードの集合をS 7とする（ステップS 8 8）。そして、S 7の各キーワードのパスのフィールドに、トップの直下のカテゴリ“その他”を設定し、処理を終了する。こうして、キーワードテーブル6 2のすべてのキーワードにパスが設定される。

【0126】

図1 6のようなディレクトリファイル5 6が生成されると、文書の分類登録が完了する。利用者は、ディレクトリアクセス部4 4および検索部4 5を介して、ディレクトリファイル5 6の必要な情報を取得することができる。

【0127】

ディレクトリアクセス部4 4への利用者からの入力としては、表示されたリンクのクリックと検索要求の2種類が考えられる。リンクがクリックされた場合は、ディレクトリアクセス部4 4は、対応するファイル1 0 2、1 0 3の内容を表示装置1 4に表示する。利用者は、図1 6のハイパーテキストによる索引を、次のように利用することができる。

【0128】

1. パス：ハイパーテキスト全体の中における現在位置が把握できる。迷子にならないための工夫である。

2. 上位関連語：キーワードに関連する上位または広い概念のカテゴリとして、検索結果を広げるために用いられる。

【0129】

3. サブカテゴリ：キーワードの下位のカテゴリとして、検索結果を絞り込むために用いられる。

4. 連想語：キーワードとの関連性は低いが、文書を通じて繋がっているカテゴリとして、ハイパーテキストのブラウジング、ジャンプ等に自由に用いられる。

【0130】

また、検索要求に関しては、ディレクトリトップファイル101の画面上の入力窓104からのキーワード検索と、ディレクトリ中間ファイル103の画面上の入力窓112からの文書内容検索の2種類がある。ディレクトリアクセス部44は、これらの検索を検索部45に指示し、その結果を受け取って表示装置14に表示する。

【0131】

キーワード検索の場合は、検索部45は、キーワードテーブル62を検索し、検索要求を満たすキーワードをリストアップする。そして、ディレクトリアクセス部44は、各キーワードのディレクトリ中間ファイル103へのリンクを、ディレクトリトップファイル101に付加する。文書内容検索の場合は、検索部45は、文書データ21の文書のうち、検索要求を満たす文書をリストアップし、ディレクトリアクセス部44は、それらのタイトルおよび内容のリストを表示する。

【0132】

例えば、利用者が“自動車のワックス”についての情報を探したいと思った場合、まず、“自動車”でキーワード検索を行って文書を絞り込み、次に、“ワックス”で文書内容を検索することで、検索結果のゴミを減らすことができる。

【0133】

単に“ワックス”を含む文書を検索しただけでは、床のワックスやスキーのワックスに関する文書等の不要な文書まで検索結果に含まれてしまう。また、“自動車”のような一般的な語は、往々にして、自動車関係の文書にはそのままの形で出現しないことが多いため、“自動車 AND ワックス”のような検索式でブール検索を行っても、良い結果が得られない場合も多い。

【0134】

以上説明したように、文書整理装置は、文書に付加されたキーワードの統計情報、文字列としての包含関係、および辞書等を用いて人手で与えた関係を統合して、ディレクトリを自動的に構築する。統計情報だけでは分類の精度は低く、既存の分類をベースにすると汎用的だが文書の特徴をうまく表現できない。本実施形態では、両者を組み合わせることで、汎用性を保ちつつ、分類の精度を向上させている。

【0135】

ディレクトリサービスにおいては、階層関係による分類以外に、上位関連語、連想語、50音・アルファベット順索引といった多様なリンクを最短パスやサブカテゴリとともに提示することで、文書への複数のパスが提供され、利用者のアクセスが支援される。したがって、このサービスでは、分類そのものより、利用者を文書にナビゲートする多くの手段を提供することに主眼が置かれている。

【0136】

また、管理者が同意語・不要語リストを明示的に与え、文書整理装置がその情報に従ってキーワードのリンクを追加・削除することで、管理者の意向をディレクトリに反映させることができる。また、管理者がキーワード間の階層関係を明示的に与え、文書整理装置がその情報に従ってディレクトリを生成することで、ハイパーテキストのリンク関係を調整することができる。

【0137】

また、文書整理装置は、今回入力されたキーワードを前回のキーワードと比較し、新規に登録されたキーワードを強調表示するため、管理者の手間をかけずに、利用者が新しい話題を把握するための手掛かりが提供される。

【0138】

また、文書整理装置は、文書のキーワードをそのままディレクトリのカテゴリとして用いるため、従来の自動分類とは違って、最初から分類ミスが生じない。さらに、文書本文の全文検索と、ディレクトリの分類とを融合することで、話題を絞りこんで検索することが可能となる。これにより、同音異義語による検索ゴミを減らすことができる。

【0139】

また、ディレクトリに含まれるキーワードの検索と文書内容の全文検索を組み合わせることで、利用者が特定の話題における文書から細かい情報を検索することが支援される。

【0140】

次に、図20から図24までを参照しながら、図2に示した文書整理装置を利用した文書整理システムの実施形態について説明する。今日のネットワーク環境においては、メールやニュースといった身の回りの文書が、コンピュータ上の文書フォルダに溜まっていることが多い。文書整理システムは、そのような文書群に対して本発明を応用し、文書整理を行う。

【0141】

図20は、このような文書整理システムの構成図である。図20の文書整理システムは、処理装置121、二次記憶装置122、キーワード抽出装置123、および利用者端末124を備える。例えば、処理装置121は、CPUとメモリを含み、利用者端末124は、入力装置と表示装置を含む。

【0142】

二次記憶装置122は、文書群のデータを含む文書フォルダ131と管理ファイル132を格納する。管理ファイル132には、同意語および不要語の集合133と、キーワードの階層関係を表すデータ134と、ディレクトリのトップとなるキーワードの集合135が含まれている。

【0143】

キーワード抽出装置123は、文書フォルダ131の各文書の形態素解析を行って、文書を単語に分割する。そして、中頻度の単語をキーワードとして取り出し、処理装置121に入力する。低頻度の単語では文書の特徴を表していない恐れがあり、高頻度の単語では他の文書にも多く現れる可能性がある。

【0144】

処理装置121は、キーワード整形器41、キーワード関係抽出器42、ディレクトリファイル生成器43、検索部45、WWW (world wide web) サーバ141を含む。

【0145】

キーワード整形器41、キーワード関係抽出器42、およびディレクトリファイル生成器43は、文書フォルダ131および管理ファイル132のデータと、キーワード抽出装置123からのキーワードとを用いて上述したような処理を行い、図16に示したような形式のディレクトリファイル142を生成する。

【0146】

また、WWWサーバ141は、図2のディレクトリアクセス部44に対応し、利用者からの指示に応じてディレクトリファイル142にアクセスする。利用者は、端末124に搭載されたWWWブラウザ143を通じて、WWWサーバ141に対する指示を入力し、ディレクトリにアクセスする。

【0147】

図21は、端末124上に表示されるディレクトリのトップ画面を示している。トップ画面の項目“コンピュータ”、“ソフトウェア”等は、管理者がトップキーワード135により与えたカテゴリに相当する。

【0148】

図22は、利用者がトップ画面または他の画面上のキーワードをクリックすることにより表示されるディレクトリの中間画面を示している。ここでは、“ホームページ”というキーワードのページが表示されている。

【0149】

右上のパス151は、トップ画面から“ホームページ”の中間画面に至るパスを表し、この中間画面はトップ画面の下位のキーワード“WWW”の下位に位置することが分かる。また、関連語152は、“ページ”、“インターネット”等が“ホームページ”の上位語であることを表し、サブトピック（サブカテゴリ）153は、“WWWページ”、“接続事業”、“HTML”等の12個のキーワードが“ホームページ”の下位語であることを表している。

【0150】

また、文書リスト154は、“ホームページ”に関連する39個の文書のそれぞれについて、タイトル、本文（一次情報）へのリンク、更新日、および連想語のリンクを示している。例えば、最初の文書タイトル“国立天文台のWWWペー

ジ”の下括弧内に記述された“日本”、“電子メール”、および“代理投稿”が、“ホームページ”の連想語である。

【0151】

図23は、ディレクトリの50音・アルファベット順索引のトップ画面を示しており、図24は、その下の中間画面を示している。図24では、“れ”で始まるキーワードがリストアップされている。このような文書整理システムによれば、任意の文書を整理して格納することができる。

【0152】

また、本実施形態の文書整理装置は、他にも次のようなシステムに応用することができる。

(1) 情報共有ツールのビュー

本出願人による先願である「文書共有整理システム、共有文書管理装置および文書アクセス装置」（特願平8-281940）では、ネットワークを通じてグループで文書情報を共有し、特定の文書のリストとしてビューを表示することができる。このビューの1つとして、文書整理装置が作成するディレクトリを表示することが考えられる。

(2) ネットワークニュース検索システム

本出願人による先願である「関連文書表示装置」（特願平10-82270）では、ネットワークニュースの検索システムが開示されている。ネットワークニュースにおけるニュースグループの整理に文書整理装置を応用することで、利用者のアクセス支援ができる。

【0153】

なお、以上説明した実施形態においては、管理者が管理ファイルを作成しているが、利用者自身が管理者の役割を兼ねることもあり得る。また、本発明は、文書のみならず、キーワードを付加されたあらゆる情報の分類・整理に応用することが可能である。例えば、画像や音声のファイルに適切なキーワードを付加しておき、それらの間の関係からディレクトリファイルを作成することができる。

【0154】

ところで、図2の文書整理装置は、図25に示すような情報処理装置（コンピ

ュータ)を用いて構成することができる。図25の情報処理装置は、CPU161、メモリ162、入力装置163、出力装置164、外部記憶装置165、媒体駆動装置166、およびネットワーク接続装置167を備え、それらはバス168により互いに接続されている。

【0155】

メモリ162は、例えば、ROM (read only memory)、RAM (random access memory) 等を含み、処理に用いられるプログラムとデータを格納する。CPU161は、メモリ162を利用してプログラムを実行することにより、必要な処理を行う。

【0156】

図2のキーワード整形器41、キーワード関係抽出器42、ディレクトリファイル生成器43、ディレクトリアクセス部44、および検索部45は、それぞれ、メモリ162の特定のプログラムコードセグメントにプログラムとして格納される。

【0157】

入力装置163は、例えば、キーボード、ポインティングデバイス、タッチパネル等であり、ユーザからの指示や情報の入力に用いられる。出力装置164は、例えば、ディスプレイやプリンタ等であり、利用者への問い合わせ、処理結果等の出力に用いられる。

【0158】

外部記憶装置165は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク (magneto-optical disk) 装置等であり、図2の二次記憶装置12として用いられる。この外部記憶装置165に、上述のプログラムとデータを保存しておき、必要に応じて、それらをメモリ162にロードして使用することもできる。

【0159】

媒体駆動装置166は、可搬記録媒体169を駆動し、その記録内容にアクセスする。可搬記録媒体169としては、メモ리카ード、フロッピーディスク、CD-ROM (compact disk read only memory)、光ディスク、光磁気ディスク

等、任意のコンピュータ読み取り可能な記録媒体が用いられる。この可搬記録媒体 169 に上述のプログラムとデータを格納しておき、必要に応じて、それらをメモリ 162 にロードして使用することもできる。

【0160】

ネットワーク接続装置 167 は、LAN (local area network) 等の任意のネットワーク (回線) を介して外部の装置と通信し、通信に伴うデータ変換を行う。また、必要に応じて、上述のプログラムとデータを外部の装置から受け取り、それらをメモリ 162 にロードして使用することもできる。

【0161】

図 26 は、図 25 の情報処理装置にプログラムとデータを供給することのできるコンピュータ読み取り可能な記録媒体を示している。可搬記録媒体 169 や外部のデータベース 170 に保存されたプログラムとデータは、メモリ 162 にロードされる。そして、CPU 161 は、そのデータを用いてそのプログラムを実行し、必要な処理を行う。

【0162】

【発明の効果】

本発明によれば、情報処理装置に蓄えられた大量の文書群を、外部から与えられたカテゴリと自動的に抽出された文書群の特徴を併用して、高い精度で自動的に分類することができる。また、分類結果に従って、文書への多様なリンクを有するディレクトリが自動的に生成され、利用者によるアクセスが支援される。

【図面の簡単な説明】

【図 1】

本発明の文書整理装置の原理図である。

【図 2】

文書整理装置の構成図である。

【図 3】

管理ファイルのデータ構造を示す図である。

【図 4】

文書単語対テーブルとキーワードテーブルを示す図である。

【図 5】

文書メタ情報のデータ構造を示す図である。

【図 6】

キーワード整形器の構成図である。

【図 7】

キーワード整形器の処理のフローチャートである。

【図 8】

キーワード関係抽出器の構成図である。

【図 9】

キーワード関係抽出器の処理のフローチャートである。

【図 10】

相関ルール抽出／評価処理のフローチャートである。

【図 11】

ルール分割を示す図である。

【図 12】

第 1 の文書集合の関係を示す図である。

【図 13】

第 2 の文書集合の関係を示す図である。

【図 14】

第 3 の文書集合の関係を示す図である。

【図 15】

第 4 の文書集合の関係を示す図である。

【図 16】

ディレクトリファイルを示す図である。

【図 17】

ディレクトリファイル生成器の処理のフローチャートである。

【図 18】

パス生成処理のフローチャート（その 1）である。

【図 19】

パス生成処理のフローチャート（その2）である。

【図 20】

文書整理システムの構成図である。

【図 21】

文書ディレクトリのトップ画面を示す図である。

【図 22】

文書ディレクトリの中間画面を示す図である。

【図 23】

文書 50 音・アルファベット順索引のトップ画面を示す図である。

【図 24】

文書 50 音・アルファベット順索引の中間画面を示す図である。

【図 25】

情報処理装置の構成図である。

【図 26】

記録媒体を示す図である。

【符号の説明】

- 1 関係抽出手段
- 2 生成手段
- 3 出力手段
- 4、32、53、134 階層関係
- 5、55 連想関係
- 11、121 処理装置
- 12、122 二次記憶装置
- 13 入力装置
- 14 表示装置
- 21 文書データ
- 22、132 管理ファイル
- 31、133 同意語／不要語
- 33、135 トップキーワード

- 4 1 キーワード整形器
- 4 2 キーワード関係抽出器
- 4 3 ディレクトリファイル生成器
- 4 4 ディレクトリアクセス部
- 4 5 検索部
- 5 1 文書単語対
- 5 2 文書メタ情報
- 5 4 同値関係
- 5 6、1 4 2 ディレクトリファイル
- 6 1 文書単語対テーブル
- 6 2 キーワードテーブル
- 7 1 キーワード統一器
- 7 2 単語読み付加器
- 7 3 旧キーワード
- 8 1 部分文字列切り出し器
- 8 2 相関ルール抽出器
- 8 3 ルール評価器
- 8 4 マージャ
- 8 5 部分語関係
- 8 6 相関ルール
- 9 1、9 2、9 3、9 4 領域
- 1 0 1 ディレクトリトップファイル
- 1 0 2 50音・アルファベット順索引中間ファイル
- 1 0 3 ディレクトリ中間ファイル
- 1 0 4、1 1 2 検索窓
- 1 0 5 キーワード
- 1 0 6 50音索引
- 1 0 7 ヘッダ
- 1 0 8、1 5 1 パス

- 109、152 上位関連語
- 110、153 サブカテゴリ
- 111、154 文書リスト
- 123 キーワード抽出装置
- 124 利用者端末
- 131 文書フォルダ
- 141 WWWサーバ
- 143 WWWブラウザ
- 161 CPU
- 162 メモリ
- 163 入力装置
- 164 出力装置
- 165 外部記憶装置
- 166 媒体駆動装置
- 167 ネットワーク接続装置
- 168 バス
- 169 可搬記録媒体
- 170 データベース

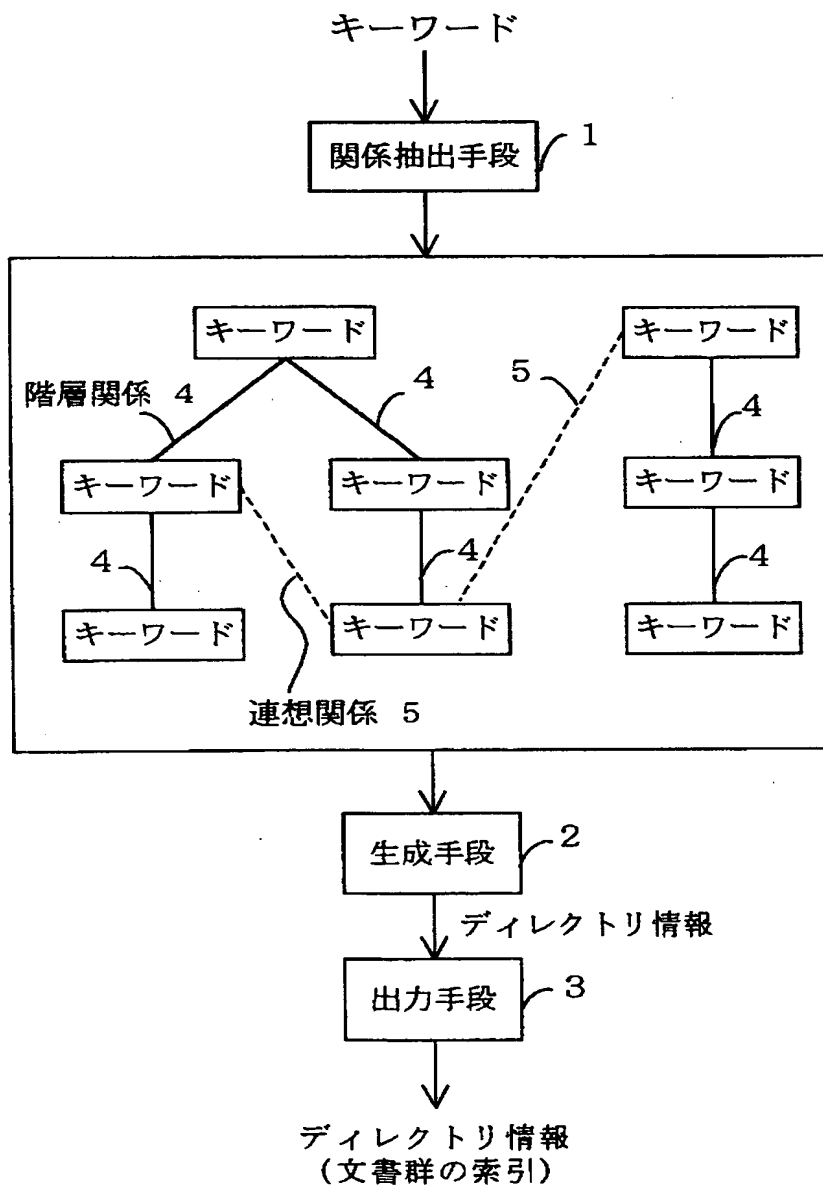
特平 10-176749

【書類名】

図面

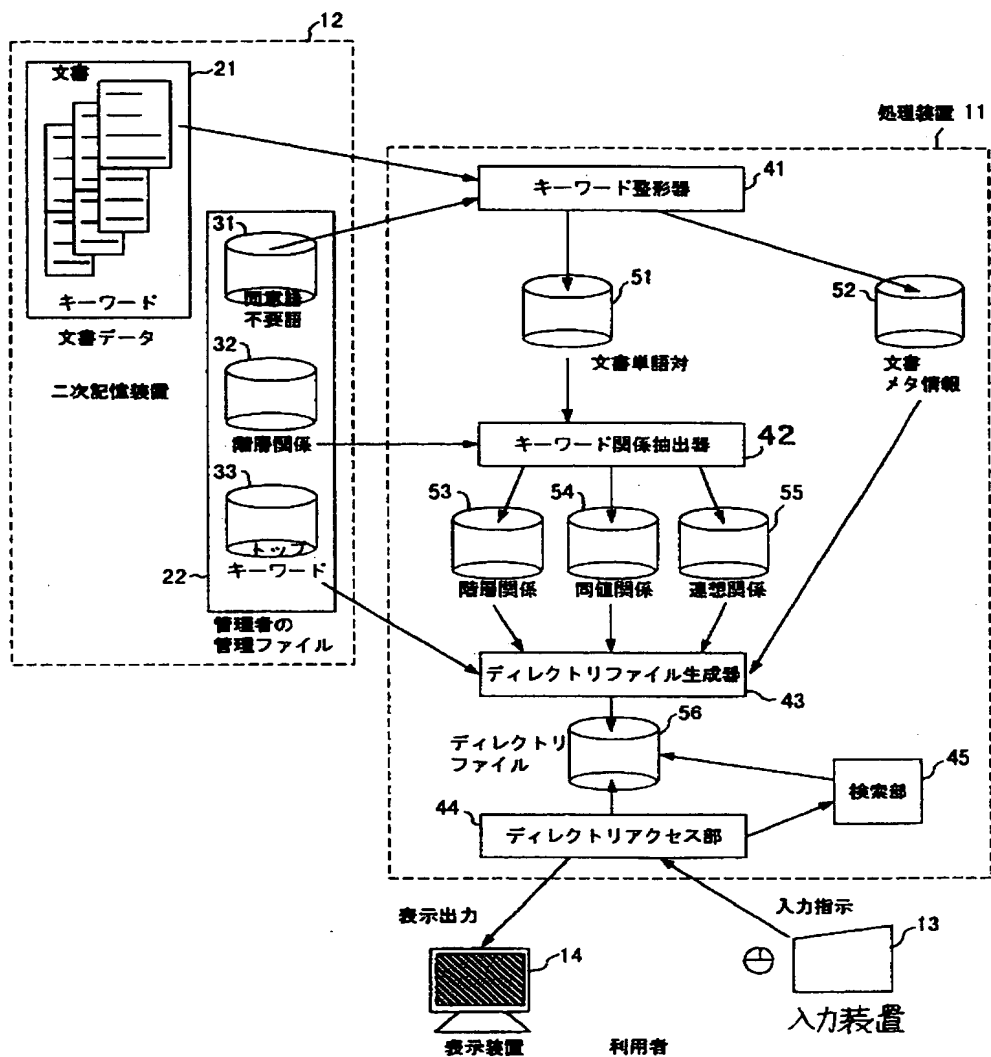
【図1】

本 発 明 の 原 理 図



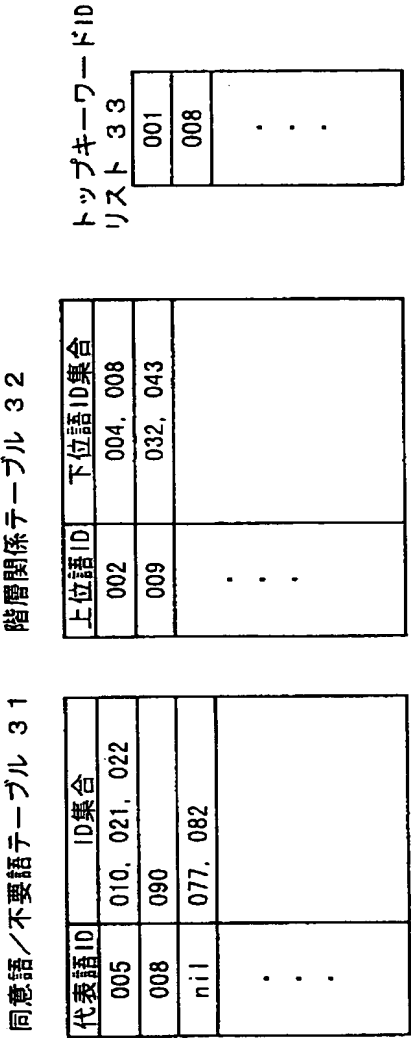
【图 2】

文書整理装置の構成図



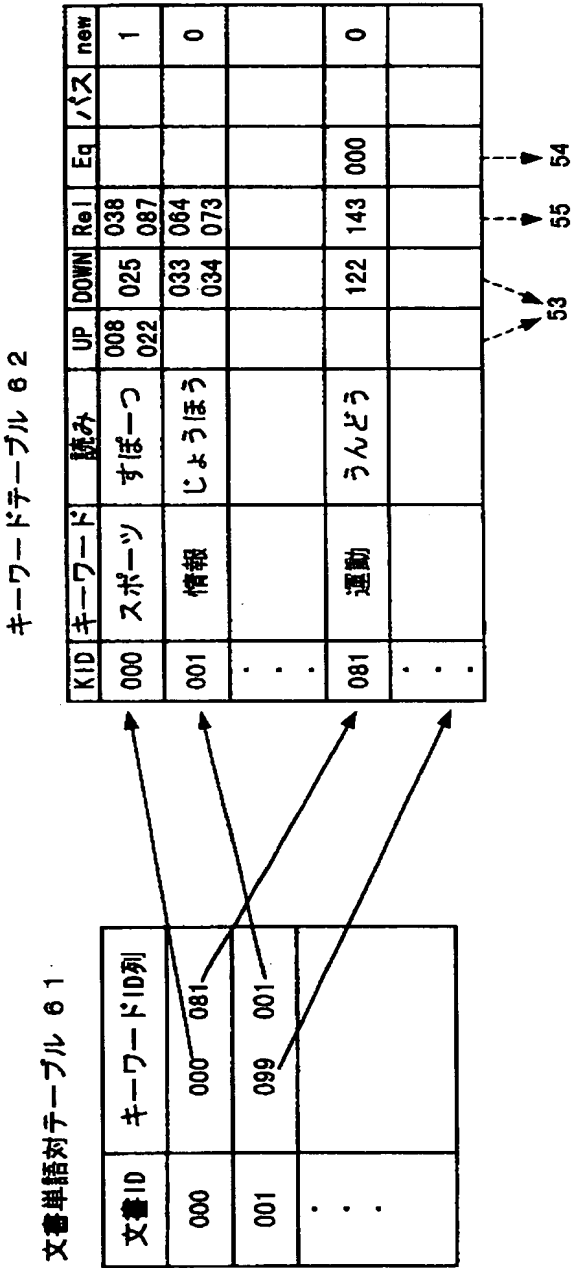
【図 3】

管理ファイルのデータ構造を示す図



【図 4】

文書単語対テーブルとキーワードテーブルを示す図



【図 5】

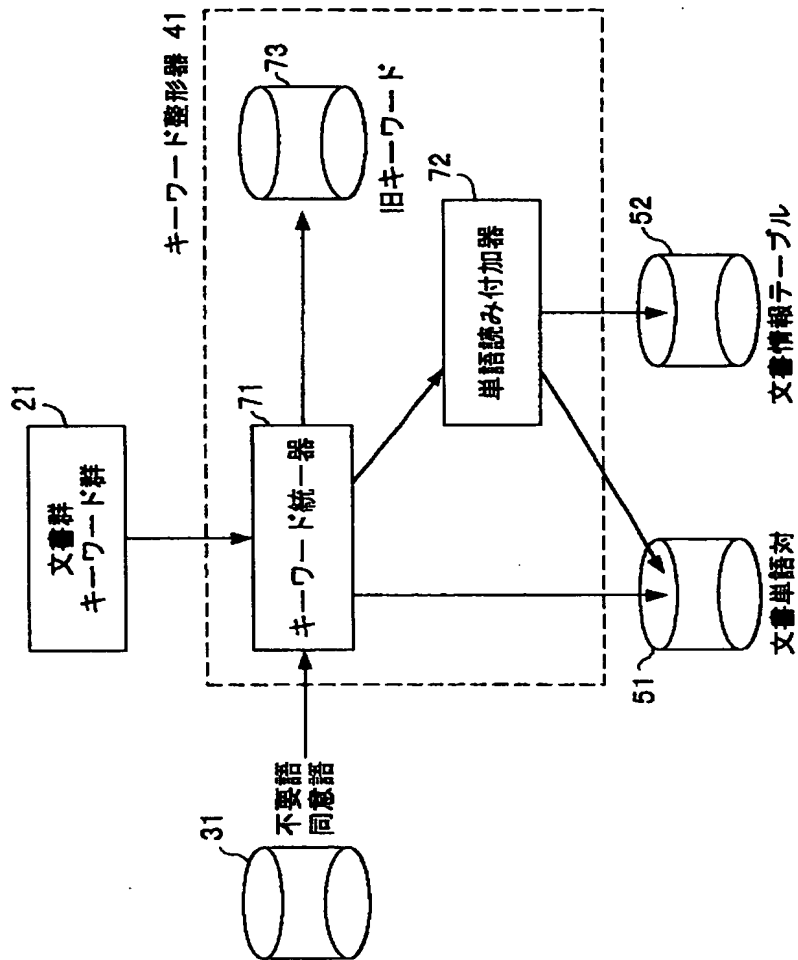
文書メタ情報のデータ構造を示す図

文書情報テーブル 52

文書ID	タイトル	説明	更新日時	一時情報リンク
001	ニュースの 読み方	comp.fjを muleから...	1998/2/10 15:38	http://www.xxxx
002				
...				

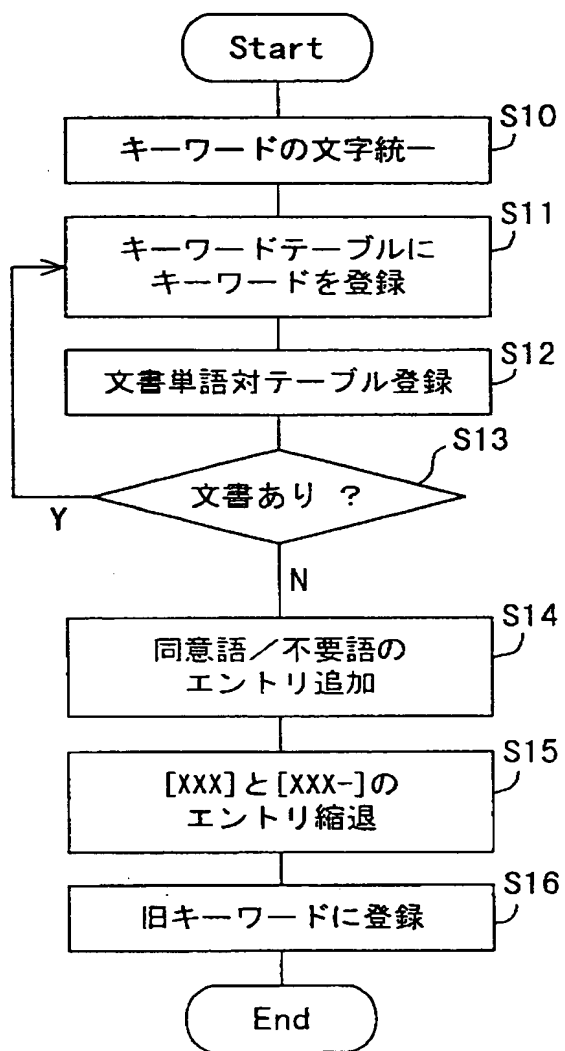
【図 6】

キーワード整形器の構成図



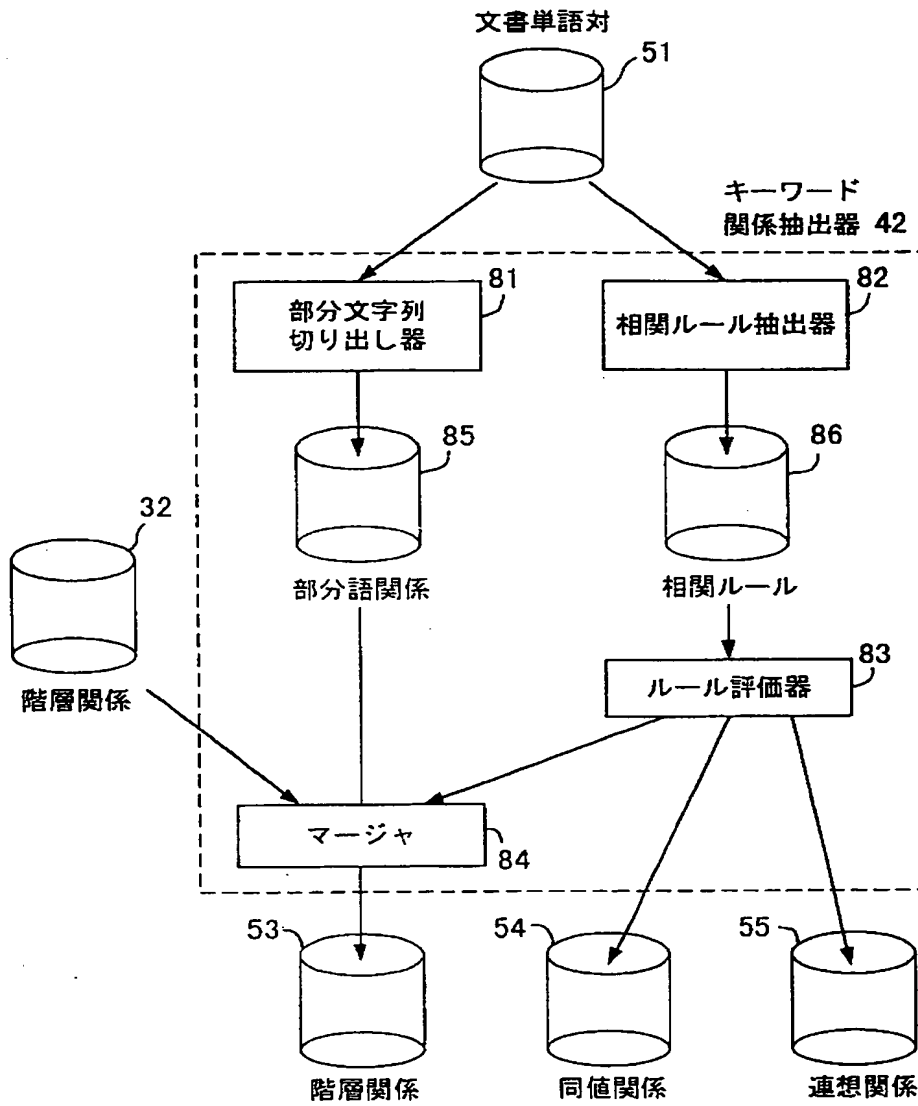
【図 7】

キーワード整形器の処理のフローチャート



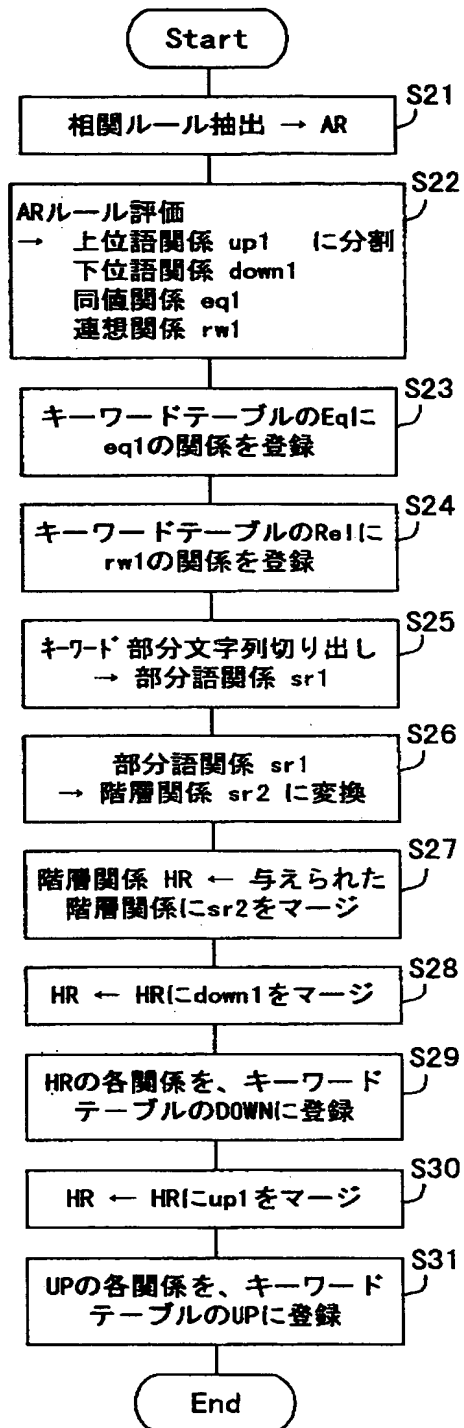
【図 8】

キーワード関係抽出器の構成図



【図 9】

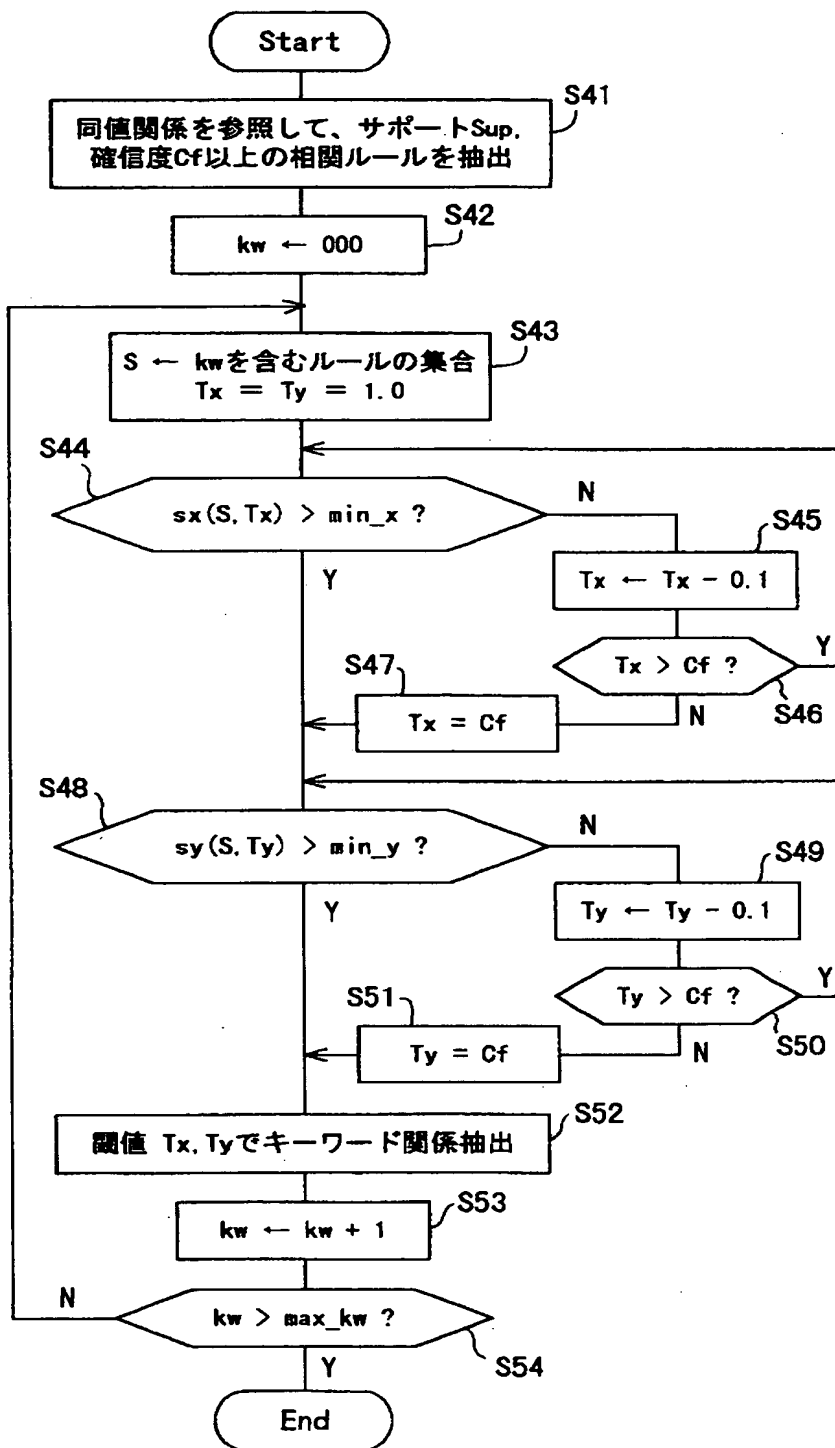
キーワード関係抽出器の処理のフローチャート



特平 10-176749

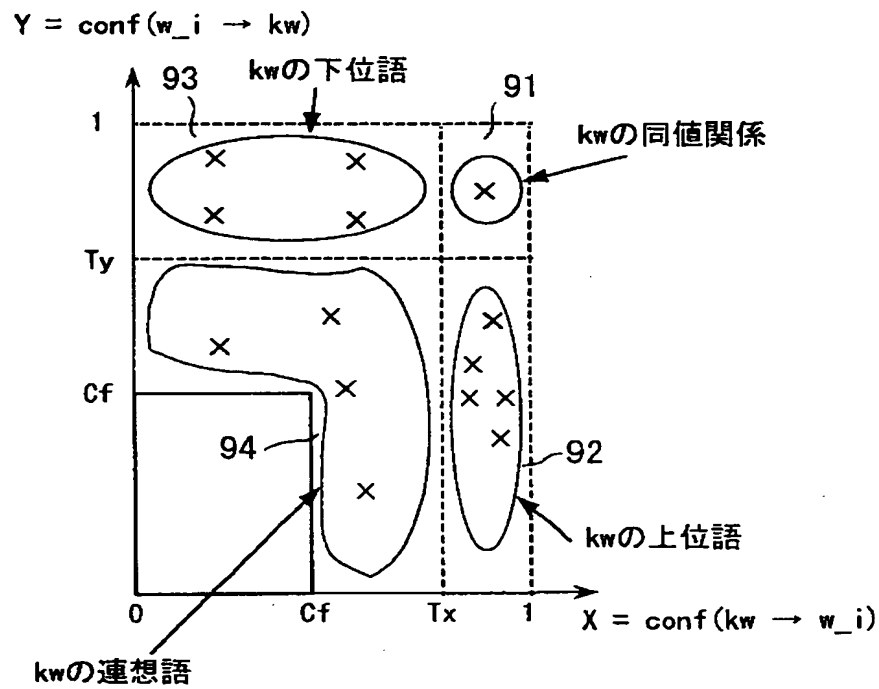
【図 10】

相関ルール抽出／評価処理のフローチャート



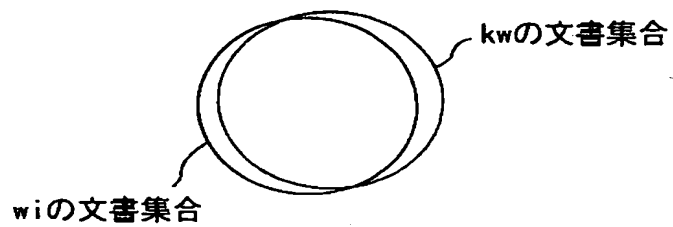
【図 11】

ルール分割を示す図



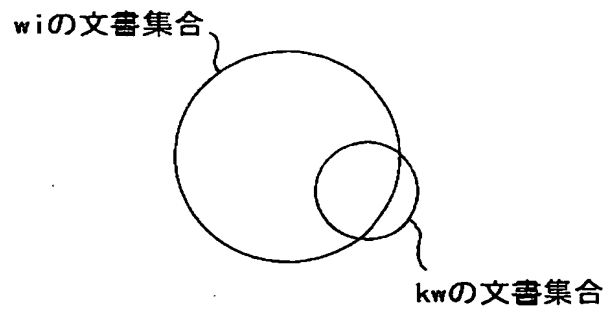
【図 12】

第 1 の文書集合の関係を示す図



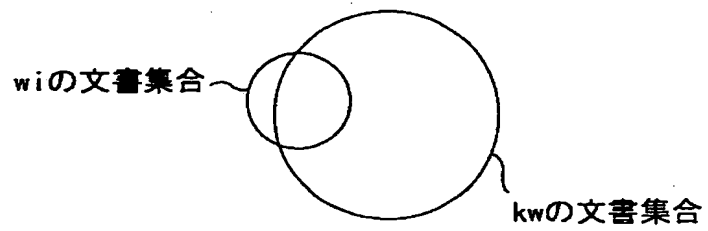
【図 13】

第 2 の文書集合の関係を示す図



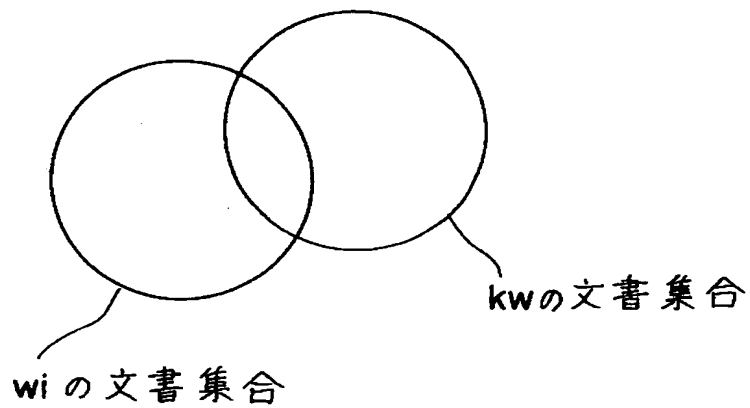
【図14】

第3の文書集合の関係を示す図



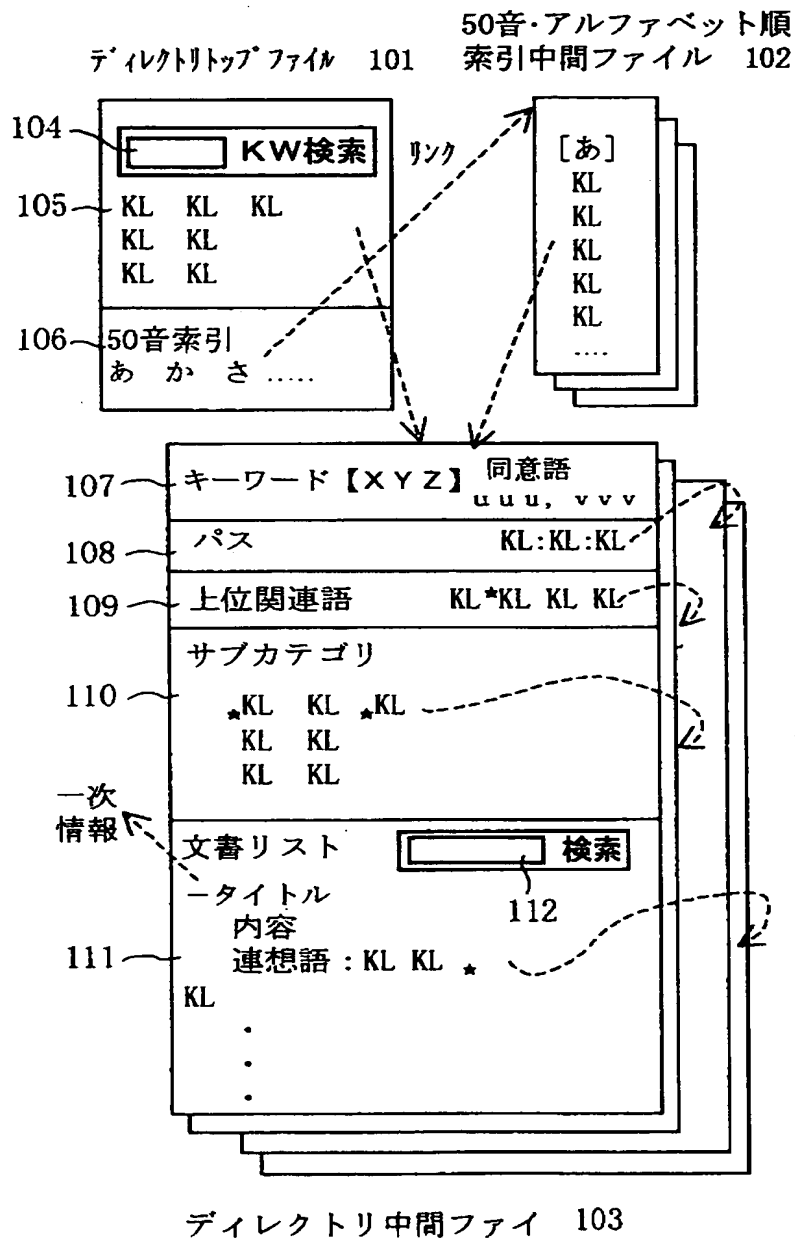
【図15】

第4の文書集合の関係を示す図



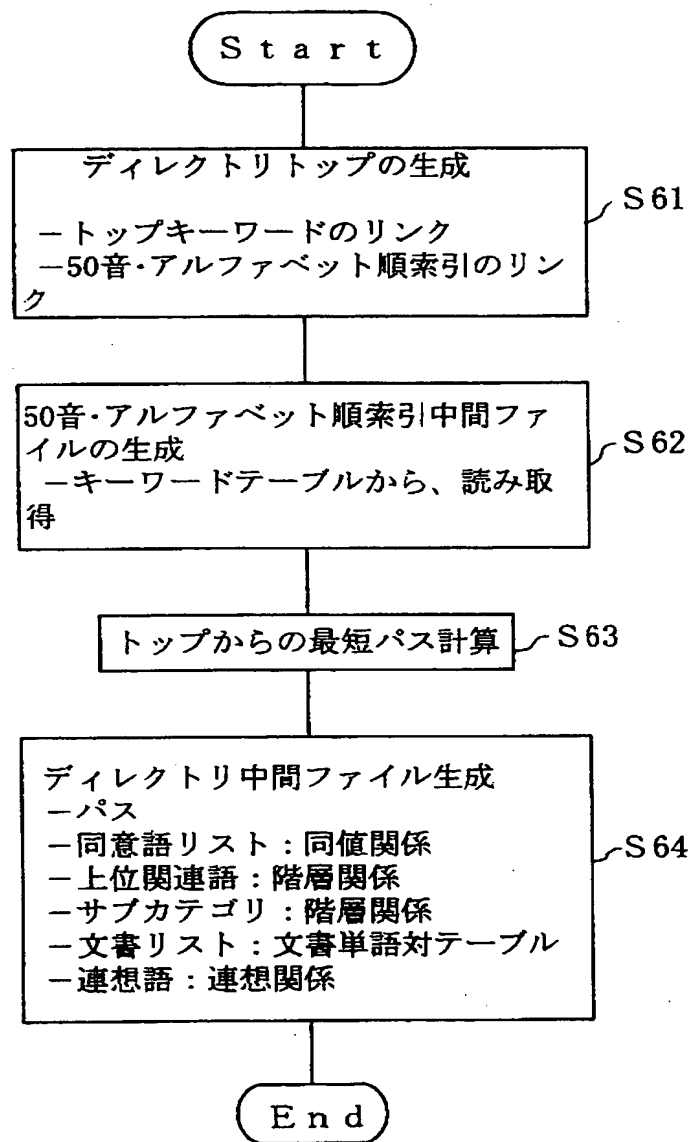
【図 16】

ディレクトリファイルを示す図



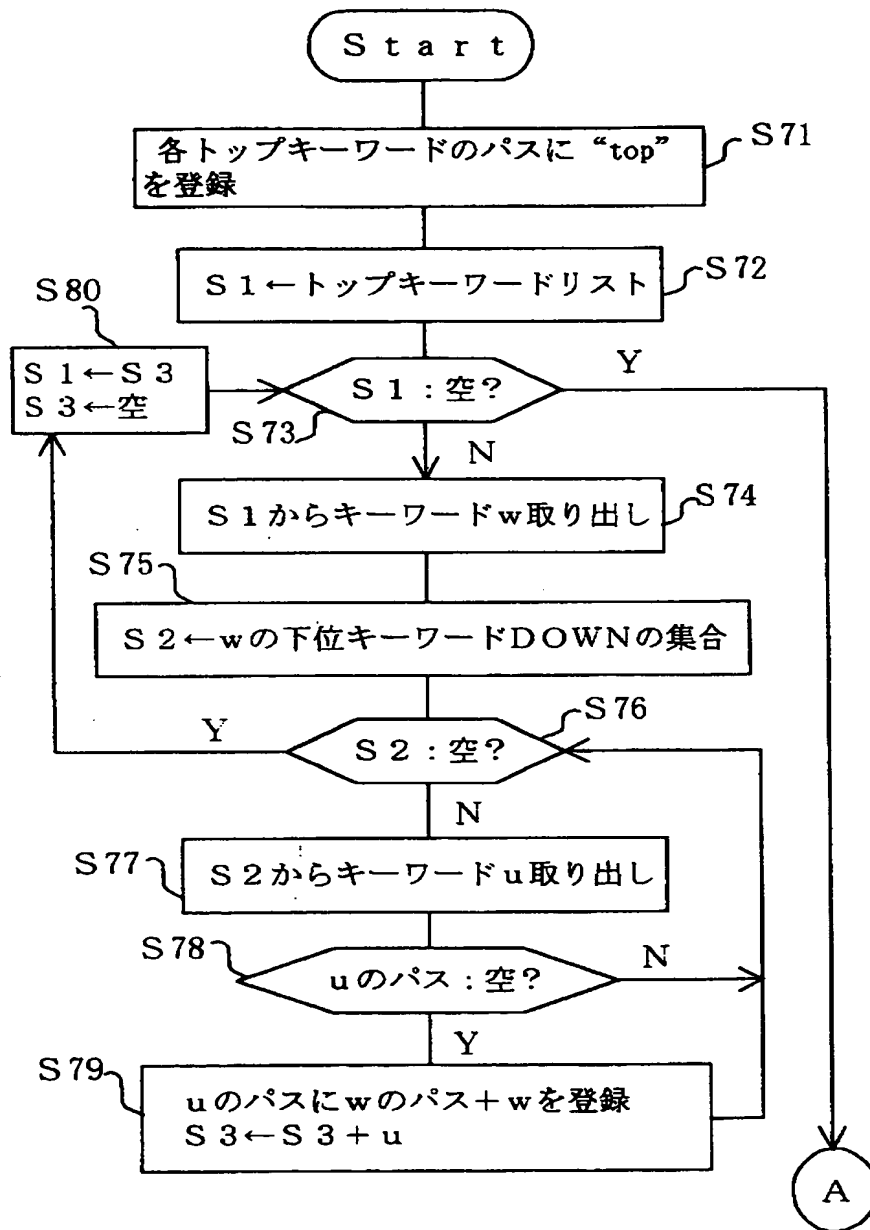
【図 17】

ディレクトリファイル生成器の
処理のフローチャート



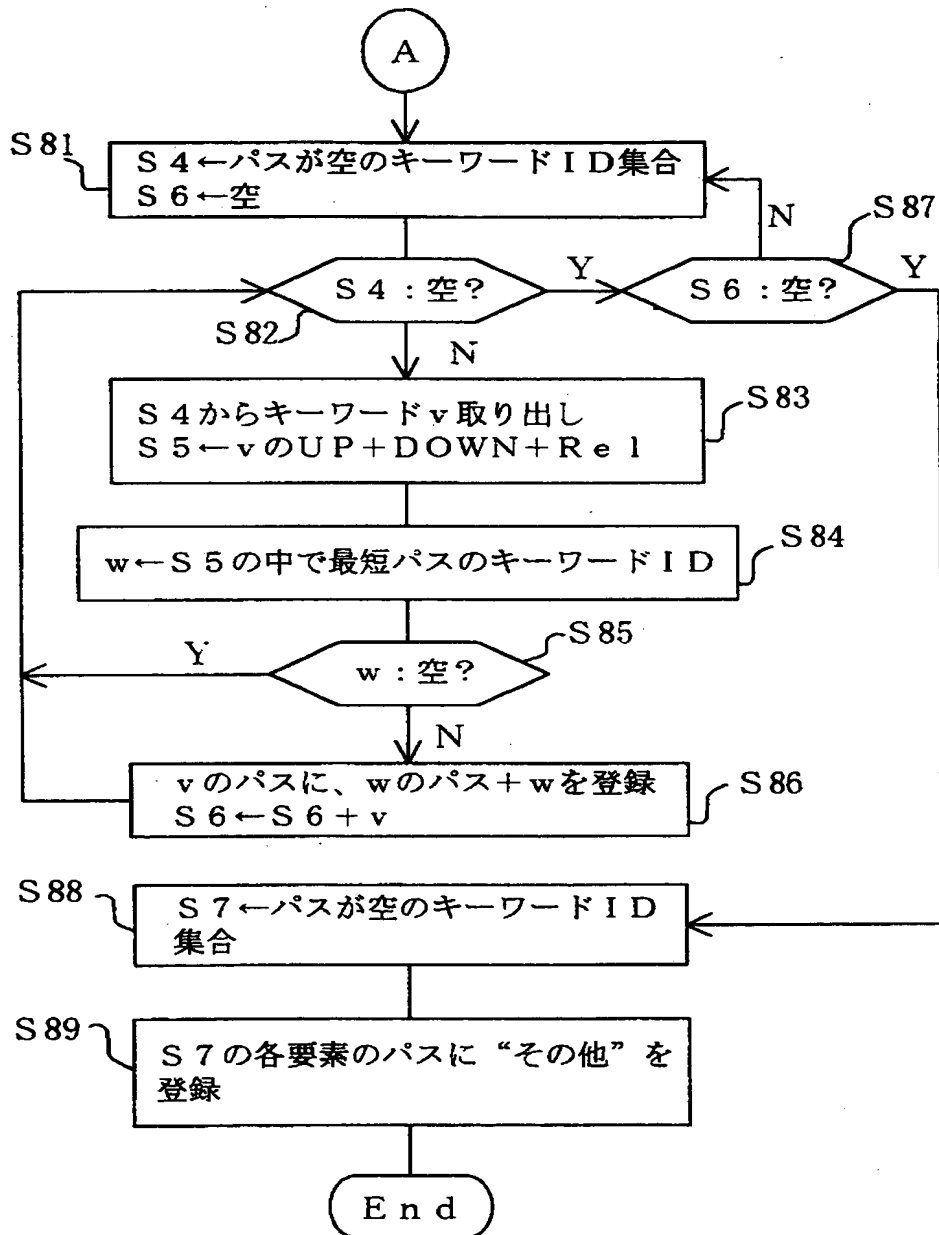
【図18】

パス生成処理のフローチャート（その1）



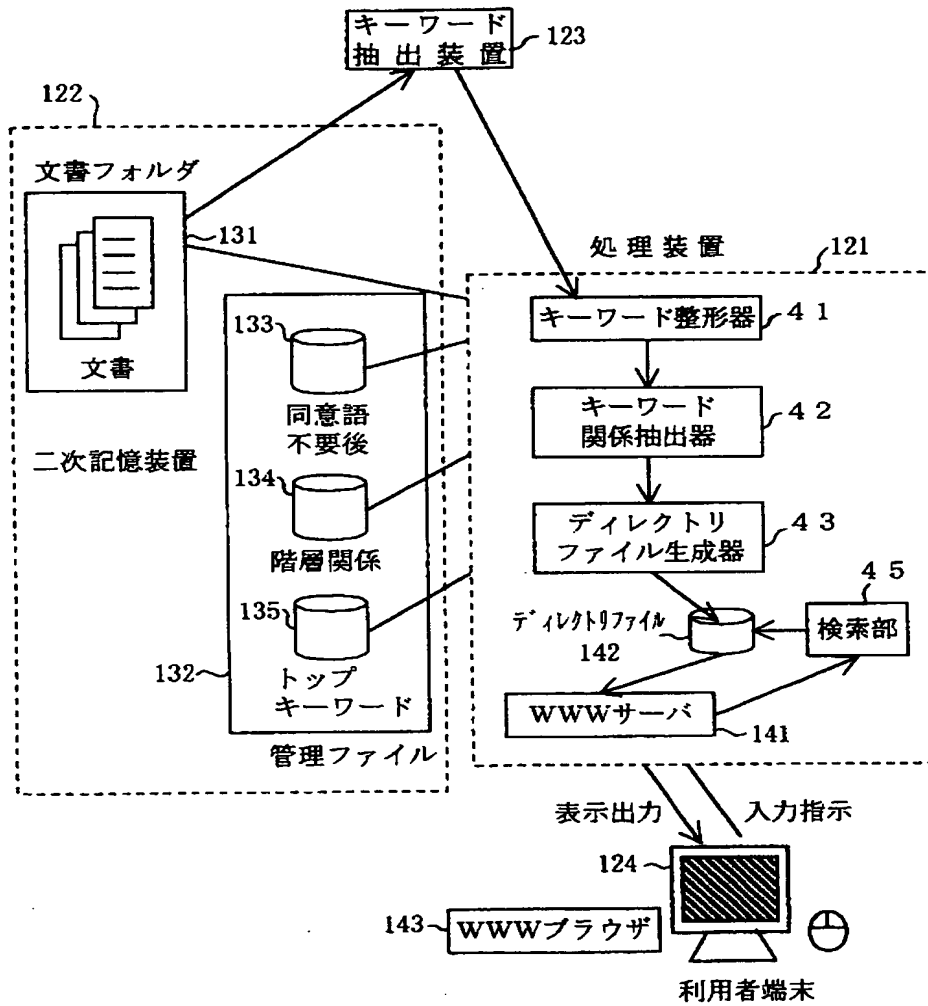
【図19】

パス生成処理のフローチャート（その2）



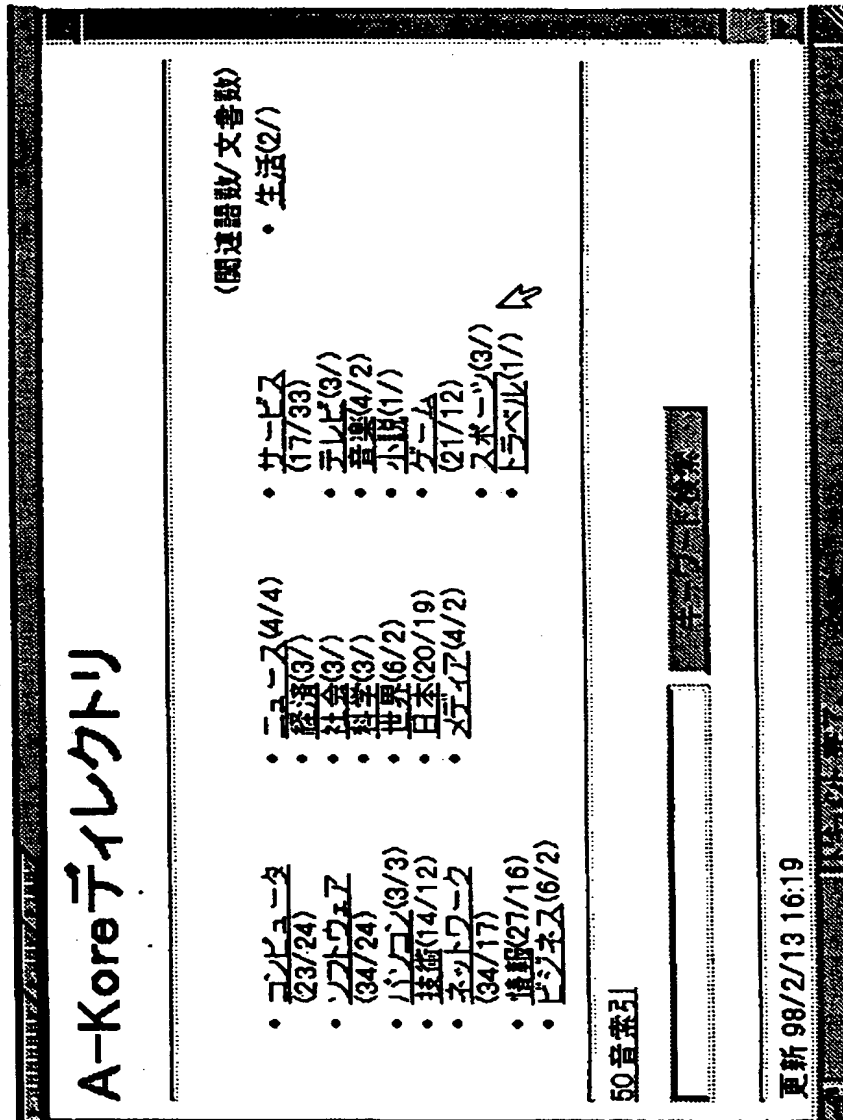
【図20】

文書整理システムの構成図



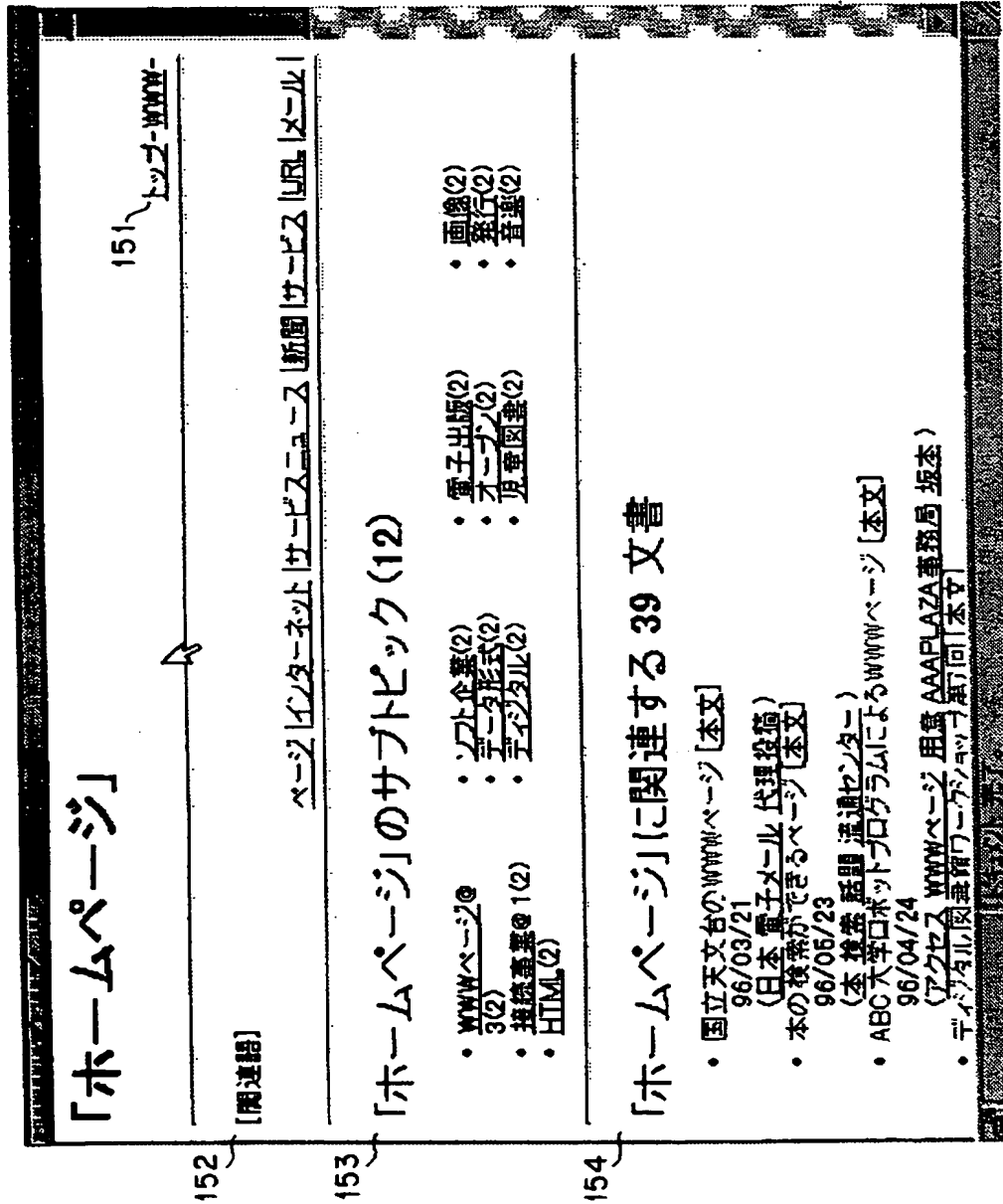
【図21】

文書ディレクトリのトップ画面を示す図



【図22】

文書ディレクトリの中画面を示す図



特平 10-176749

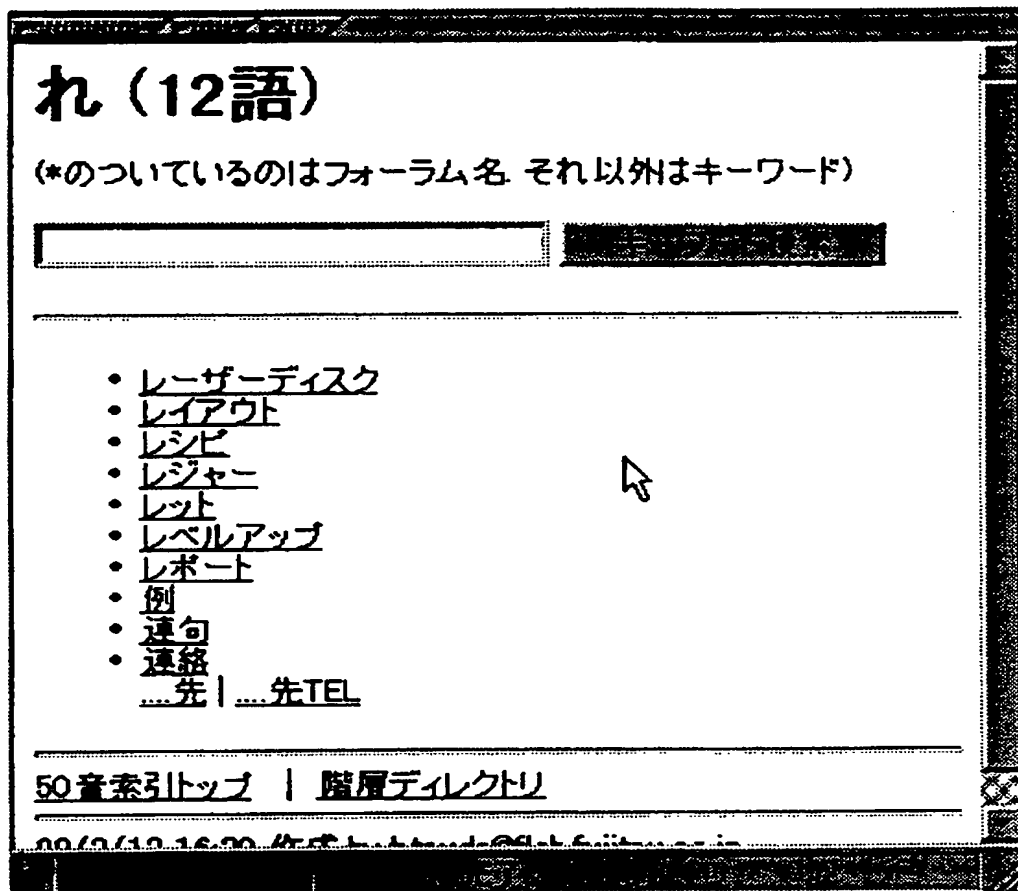
【図 23】

文書50音・アルファベット順索引のトップ画面
を示す図

A-Kore五十音索引	
	A
(数字はキーワード数)	
あ	あ20 い25 え26 お11 か15
か	か59 き15 こ24 さ7 し9 す6 せ26 じ4
さ	さ42 さ2 し45 し34 す10 せ21 ぜ6 じ3
し	し14 じ11 じ19 じ29 じ10 ぜ25 ぜ27 じ7
じ	じ11 し34 じ4 じ1
は	は14 ひ7 ふ11 ぶ6 ぶ5 ふ26 ふ3 ふ17 ふ4 ふ2 ふ17 ふ2 ふ3
ふ	ふ21 ぶ5 ふ1 ぶ17 ぶ15
ひ	ひ3 ひ10 ひ7
ふ	ふ5 ひ11 ひ1 ひ12 ひ3
ひ	ひ5
アルファベット	あ36 い13 お57 か27 き26 こ17 さ1 し9 す6 せ26 じ4 ぜ6 じ3
配列	あ36 い13 お57 か27 き26 こ17 さ1 し9 す6 せ26 じ4 ぜ6 じ3
アルファベット	あ36 い13 お57 か27 き26 こ17 さ1 し9 す6 せ26 じ4 ぜ6 じ3
配列	あ36 い13 お57 か27 き26 こ17 さ1 し9 す6 せ26 じ4 ぜ6 じ3
アルファベット	あ36 い13 お57 か27 き26 こ17 さ1 し9 す6 せ26 じ4 ぜ6 じ3
配列	あ36 い13 お57 か27 き26 こ17 さ1 し9 す6 せ26 じ4 ぜ6 じ3
A-Kore五十音索引	
no. 2/19. 1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50.	

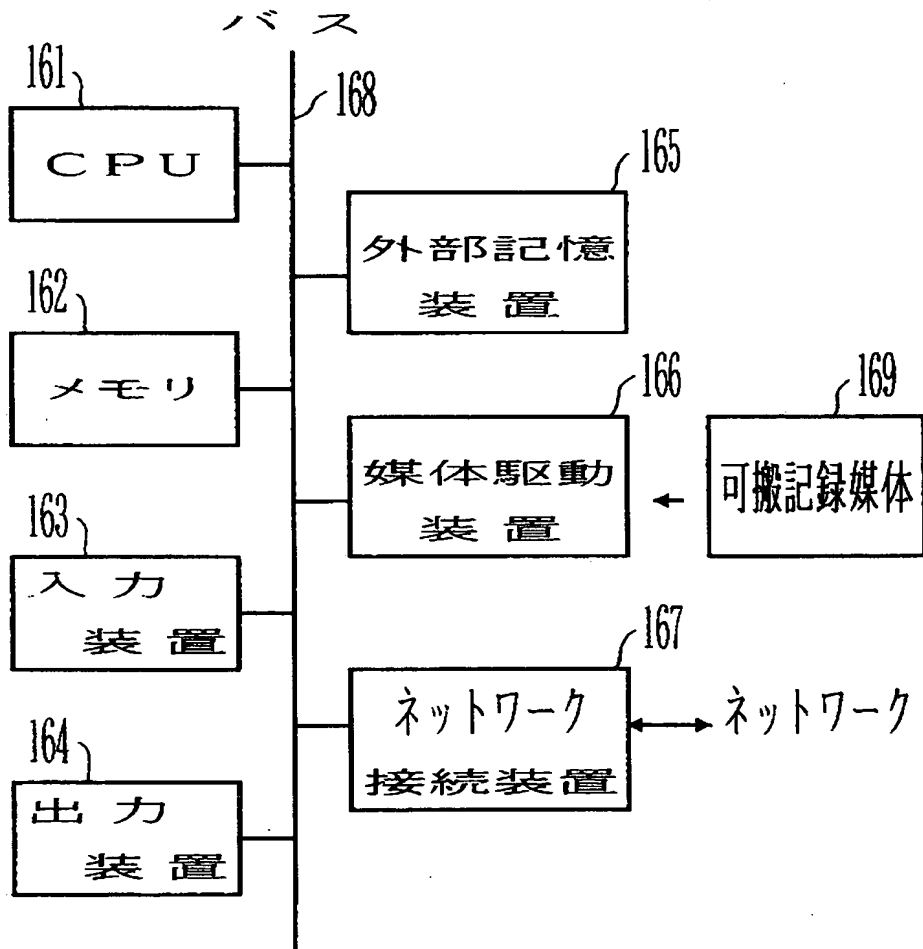
【図24】

文書50音・アルファベット順索引の中間画面を
示す図



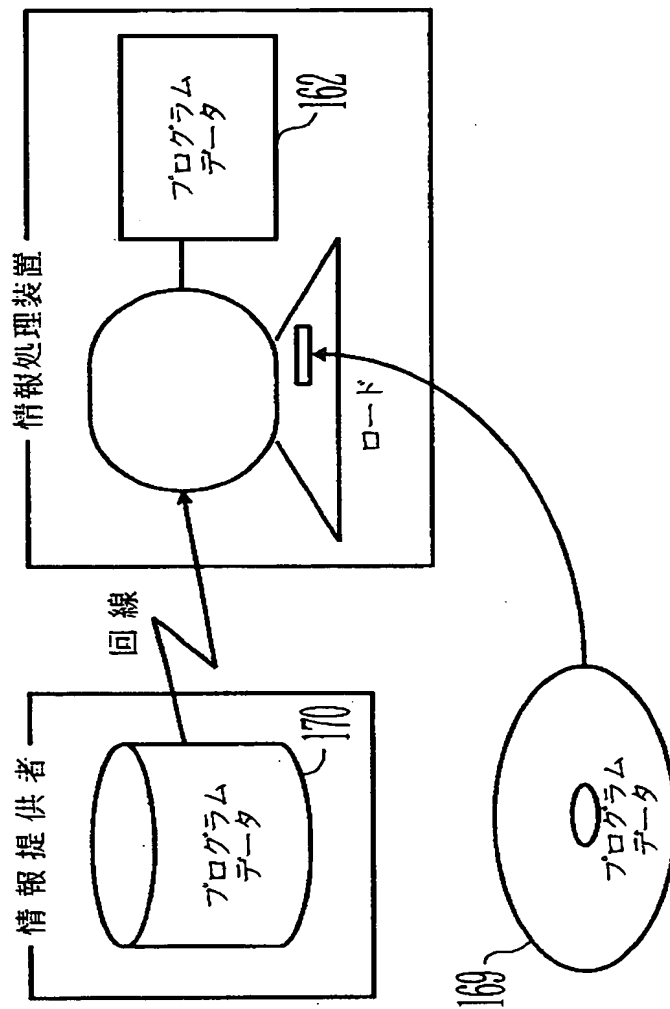
【図 25】

情報処理装置の構成図



【図 26】

記録媒体を示す図



【書類名】 要約書

【要約】

【課題】 情報処理装置に蓄えられた大量の文書群を、その特徴に従って高い精度で自動的に分類することが課題である。

【解決手段】 キーワード整形器 41 は、同意語・不要語 31 を参照して、文書データ 21 から文書単語対 51 と文書メタ情報 52 を生成する。キーワード関係抽出器 42 は、階層関係 32 を参照して、階層関係 53、同値関係 54、連想関係 55 を抽出し、ディレクトリファイル生成器 43 は、トップキーワード 33、階層関係 53、同値関係 54、連想関係 55、および文書メタ情報 52 からディレクトリファイル 56 を生成する。利用者は、ディレクトリアクセス部 44 を介してディレクトリファイル 56 にアクセスする。

【選択図】 図 2

【書類名】
【訂正書類】

職権訂正データ
特許願

<認定情報・付加情報>

【特許出願人】

【識別番号】

000005223

【住所又は居所】

神奈川県川崎市中原区上小田中4丁目1番1号

【氏名又は名称】

富士通株式会社

【代理人】

申請人

【識別番号】

100074099

【住所又は居所】

東京都千代田区二番町8番地20 二番町ビル3F
大管内外国特許事務所

【氏名又は名称】

大菅 義之

【選任した代理人】

【識別番号】

100067987

【住所又は居所】

神奈川県横浜市港北区太尾町1418-305 (大倉山二番館) 久木元特許事務所

【氏名又は名称】

久木元 彰

出 願 人 履 歴 情 報

識別番号 [000005223]

1. 変更年月日 1996年 3月26日

[変更理由] 住所変更

住 所 神奈川県川崎市中原区上小田中4丁目1番1号
氏 名 富士通株式会社